Technical Report 892

Army Synthetic Validity Project: Report of Phase II Results

Volume I

Norman G. Peterson and Cyndi Owens-Kurtz
Personnel Decision Research Institute, Inc.

R. Gene Hoffman
Human Resources Research Organization

Jane M. Arabian
U.S. Army Research Institute

Deborah L. Whetzel
American Institutes for Research

June 1990

FIFTY YEARS OF SERVICE
1940 - 1990

**United States Army Research Institute
for the Behavioral and Social Sciences**

U.S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON	**JON W. BLADES**
Technical Director	**COL, IN**
	Commanding

Research accomplished under contract for the Department of the Army

American Institutes for Research

Technical review by

Jane M. Arabian
Jacinto Manuel Silva

NOTICES

~~DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.~~

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS
--

2a. SECURITY CLASSIFICATION AUTHORITY
--

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE
--

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release;
distribution is unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
--

5. MONITORING ORGANIZATION REPORT NUMBER(S)
ARI Technical Report 892

6a. NAME OF PERFORMING ORGANIZATION
American Institutes for Research

6b. OFFICE SYMBOL *(If applicable)*
--

7a. NAME OF MONITORING ORGANIZATION
U.S. Army Research Institute for the Behavioral and Social Sciences

6c. ADDRESS *(City, State, and ZIP Code)*
3333 K Street, NW, Suite 300
Washington, DC 20007

7b. ADDRESS *(City, State, and ZIP Code)*
PERI-RS
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
U.S. Army Research Institute for the Behavioral and Social Sciences

8b. OFFICE SYMBOL *(If applicable)*
PERI-R

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
MDA903-87-C-0525

8c. ADDRESS *(City, State, and ZIP Code)*
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
62785A	791	231	C1

11. TITLE *(Include Security Classification)*
Army Synthetic Validity Project: Report of Phase II Results, Volume I

12. PERSONAL AUTHOR(S) Peterson, Norman G.; Owens-Kurtz, Cyndi (PDRII); Hoffman, R. Gene (HumRRO); Arabian, Jane M. (ARI); and Whetzel, Deborah L. (AIR)

13a. TYPE OF REPORT
Interim

13b. TIME COVERED
FROM 88/12 TO 89/09

14. DATE OF REPORT *(Year, Month, Day)*
1990, June

15. PAGE COUNT

16. SUPPLEMENTARY NOTATION Prepared in cooperation with subcontractors (Personnel Decision Research Institute, Inc. (PDRII) and Human Resources Research Organization (HumRRO). Contracting Officer's Representative, Jane M. Arabian.

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Synthetic validation
Standard setting
Job description
Expert judgment
Project A

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The two major objectives of the Army Synthetic Validity Project are to identify and evaluate procedures for

- identifying an optimal composite of selection measures for any Army enlisted Military Occupational Specialty (MOS) and estimating the validity of this composite for predicting job performance, and

- setting a minimum qualifying score to assure a reasonable probability of successful job performance, as well as other appropriate cutting scores for other critical selection decisions (e.g., for selecting recruits with potential for outstanding performance).

(Continued)

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Jane M. Arabian

22b. TELEPHONE *(Include Area Code)*
(703) 274-8275

22c. OFFICE SYMBOL
PERI-RS

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

ARI Technical Report 892

19. ABSTRACT (Continued)

Synthetic validation approaches typically begin with identification of job components that can be used to describe the population of jobs being studied. A prediction equation is derived for linking available selection tests to each component. Subject matter experts (SMEs) are asked to identify the importance of each component when compared to overall job performance. Finally, the prediction equations for the various components are weighted according to the importance judgment weights and summed to obtain an equation for predicting overall performance for the job.

The standard-setting task of the Synthetic Validity Project requires development of procedures for specifying minimum qualifying scores and other appropriate cut scores on the predictor composites identified for each job. Procedures will be developed for identifying job performance standards for each job, and these performance standards will then be linked to scores on the predictor composite for that job.

There are three research phases in the Project. In each phase, synthetic validation procedures and standard-setting procedures are developed or refined and then tried out on a new sample of MOS.

A major goal in Phase II for synthetic validation was to replicate and to extend Phase I procedures for generating synthetic prediction equations for seven MOS. Four job component models (consisting of tasks, activities, attributes, and a hybrid of tasks and activities) were used to obtain job description judgments. Predictors were linked via expert judgment to the job components. Various ways of generating prediction equations were investigated. A second goal was to evaluate differences in the job descriptions generated by different types of judges.

A major goal in Phase II standard setting was to refine the three different methods for setting performance standards. Three standard-setting methods reflecting performance on tasks, critical incidents, and by soldiers were used to obtain component standards. Again, we collected judgments for combining the component standards.

In the final phase of the project, we will refine the task and critical incident methods to yield better agreement among judges and greater convergence across methods.

Technical Report 892

Army Synthetic Validity Project: Report of Phase II Results

Volume I

Norman G. Peterson and Cyndi Owens-Kurtz
Personnel Decision Research Institute, Inc.

R. Gene Hoffman
Human Resources Research Organization

Jane M. Arabian
U.S. Army Research Institute

Deborah L. Whetzel
American Institute for Research

Selection and Classification Technical Area
Michael G. Rumsey, Chief

Manpower and Personnel Research Laboratory
Paul A. Gade, Acting Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

June 1990

Army Project Number
2Q162785A791

Manpower, Personnel, and Training

FOREWORD

In 1980 the Assistant Secretary of Defense directed all services to pursue a long-range systematic program to validate the Armed Services Vocational Aptitude Battery (ASVAB) and to re-evaluate enlistment standards against on-the-job performance. The Army has oeen investigating the validity of the ASVAB, as well as several new predictor measures, for a sample of 20 diverse MOS. This effort, known as Project A, has been very successful in validating the ASVAB and providing the Army with a greater understanding of knowledge, skills, abilities, and other personal characteristics (KSAOs) required for these 20 MOS.

A major question now facing the Army is how to extend the wealth of data collected for Project A to the other 250-plus entry-level Army MOS and to new MOS created for new hardware systems as they become operational. A second challenge is to determine the methods needed for setting job performance standards that can be used in making selection and classification decisions.

The Army currently has a research project, the Synthetic Validity Project (SYNVAL), that addresses these challenges. Specifically, the objectives of SYNVAL are to (1) evaluate synthetic validation techniques for determining MOS-specific selection composites for each MOS, and (2) evaluate alternative methods for setting minimum qualifying scores on each of these composites. The research will proceed in three phases. Phase II was recently completed and this document provides information on Phase II research plans, objectives, and results.

Based on the results of the evaluations, recommendations will be made for the most promising approach for (1) a method for developing job performance prediction equations for all of the Army's 250-plus MOS, and (2) a method for setting performance standards for these MOS. The technical quality of this project is guided by the Scientific Advisory Committee: Drs. Phil Bobko (Chair), Robert Linn, Richard Jaeger, Joyce Shields, and Robert Guion.

Edgar M Johnson

EDGAR M. JOHNSON
Technical Director

ARMY SYNTHETIC VALIDITY PROJECT: REPORT OF PHASE II RESULTS VOLUME I

EXECUTIVE SUMMARY

Requirement:

The two major objectives of the Army Synthetic Validity Project are to identify and evaluate procedures for

- identifying an optimal composite of selection measures for any Army enlisted Military Occupational Specialty (MOS) and estimating the validity of this composite for predicting job performance, and

- setting a minimum qualifying score to assure a reasonable probability of successful job performance, as well as other appropriate cutting scores for other critical selection decisions (e.g., for selecting recruits with potential for outstanding performance).

Synthetic validation approaches typically begin with identification of job components that can be used to describe the population of jobs being studied. A prediction equation is derived for linking available selection tests to each component. Subject matter experts (SMEs) are asked to identify the importance of each component when compared to overall job performance. Finally, the prediction equations for the various components are weighted according to the importance judgment weights and summed to obtain an equation for predicting overall performance for the job.

The standard-setting task of the Synthetic Validity Project is charged with developing procedures for specifying minimum qualifying scores and other appropriate cut scores on the predictor composites identified for each job. Procedures will be developed for identifying job performance standards for each job, and these performance standards will then be linked to scores on the predictor composite for that job.

Procedure:

There are three research phases in the Project. In each phase, synthetic validation procedures and standard-setting procedures are developed or refined and then tried out on a new sample of MOS.

A major goal in Phase II for synthetic validation was to replicate and to extend Phase I procedures for generating synthetic prediction equations for seven MOS. Four job component models (consisting of tasks, activities, attributes, and a hybrid of tasks and activities) were used to obtain job description judgments. Predictors were linked via expert judgment to the job components. Various ways of generating prediction equations were investigated. A second goal was to evaluate differences in the job descriptions generated by different types of judges.

A major goal in Phase II standard setting was to refine the three different methods for setting performance standards. Three standard-setting methods reflecting performance on tasks, critical incidents, and by soldiers were used to obtain component standards. Again, we collected judgments for combining the component standards.

Findings:

For synthetic validation, Phase II results replicated Phase I results very well. Each of the four job component models produced reliable and comprehensive job descriptions. Using job description and job component validity information gathered in Phase I, we formed prediction equations that had high predictive validity for each of the three jobs. Based on the job description reliability and synthetic validity results, we recommended that the task job description be retained with modest revisions.

For standard setting, the three methods for setting component standards resulted in different standards and also in some differences in the degree of consensus among judges in setting the standards. In deriving an overall standard from component standards, we again found that a linear compensatory model accurately captured the judges' aggregation strategies.

Utilization of Findings:

At the conclusion of Phase II, we have shown that synthetic validation yields valid predictions for ten jobs. The final phase of the project will extend the validity to different Project A jobs and a new job. We will conduct the job description using the task model and explore different ways of generating prediction equations to yield better differential prediction among jobs.

Meaningful performance standards were obtained for the three jobs. In the final phase of the project, we will refine the task and critical incident methods to yield better agreement among judges and greater convergence across methods.

ARMY SYNTHETIC VALIDITY PROJECT: REPORT OF PHASE II RESULTS
VOLUME I

CONTENTS

Page

Page

LIST OF TABLES

CONTENTS (Continued)

Page

LIST OF FIGURES

CONTENTS (Continued)

Page

ARMY SYNTHETIC VALIDITY PROJECT: REPORT OF PHASE II RESULTS VOLUME I

CHAPTER 1: INTRODUCTION

Lauress L. Wise (American Institutes for Research) and
Norman G. Peterson (Personnel Decisions Research Institute, Inc.)

Overall Objectives

The two major objectives of the Army Synthetic Validity Project are to identify and evaluate procedures for

- identifying an optimal composite of selection measures for any Army enlisted Military Occupational Specialty (MOS) and estimating the validity of this composite for predicting job performance, and
- setting a minimum qualifying score so as to assure a reasonable probability of successful job performance, as well as other appropriate cutting scores for other critical selection decisions (e.g., for selecting recruits with potential for outstanding performance).

Synthetic validation approaches typically begin with the identification of a set of job components that can be used to describe the population of jobs being studied. A prediction equation is derived for linking available selection tests to each component. Subject matter experts (SMEs) are asked to identify the importance of each component to overall job performance. Finally, the prediction equations for the various components are weighted according to the importance judgment weights and summed to obtain an equation for predicting overall performance for the job.

The standard setting task of the Synthetic Validity Project is charged with developing procedures for specifying minimum qualifying scores and other appropriate cut scores on the predictor composites identified for each job. Procedures are being developed for identifying job performance standards for each job, and these performance standards will then be linked to scores on the predictor composite for that job.

There are three research phases in the Project. In each phase, synthetic validation procedures and standard setting procedures are developed or refined and then tried out on a new sample of MOS.

Phase I Objectives

A major goal in Phase I for synthetic validation was to obtain and evaluate synthetic prediction equations for three MOS, 11B (Infantryman), 63B (Light-wheel Vehicle Mechanic), and 71L (Administrative Specialist). Three job component models (consisting of tasks, activities, or attributes) were developed and used to obtain job description judgments. Predictors were linked via expert judgment to the job components. Various ways of generating prediction equations were investigated. A second goal was to evaluate differences in the job descriptions generated by different types of judges.

A major goal in Phase I standard setting was to investigate different ways of setting performance standards. Performance level definitions were developed. Three standard setting methods reflecting performance on tasks (Task-based), behvavioral examples (Critical Incident-based), and asking soldiers directly (Soldier-based) were developed to obtain component standards. One method was developed for combining the component standards.

Phase I Findings

For synthetic validation, the completion of Phase I represented a major accomplishment for the project. First, we have shown that synthetic validation can be successfully carried out for the three Phase I MOS. Army SMEs were able to use the three job component models to reliably describe the content of those jobs. Table 1.1 shows, for the Task Category and Job Activity instruments, adequate single-rater reliability esimates of importance ratings for Core Technical Proficiency[1] and Overall Job Performance. Table 1.1 also shows adequate reliability estimates for attribute validity ratings from soldiers and psychologists for Core Technical Proficiency. Using job description and job component validity information, we formed prediction equations that were valid[2] for predicting Core Technical Proficiency for each of the three jobs (see Table 1.2). However, as Table 1.2 also shows, the prediction equations, on average, offered little or no discriminant validity[3].

For standard setting, Army SMEs found the performance level definitions to be reasonable and workable. Many SMEs also reported that the outcomes of the performance levels were realistic. As Table 1.3 shows, the three methods for setting standards resulted in different standards. These methods also resulted in some differences in the degree of consensus among judges in setting the standards. Compared to the Critical Incident and Soldier-based methods, the Task-based method resulted in strictest standards, which meant that it had the highest proportion of unacceptable performance among incumbents. We also found that SMEs reported difficulties in providing task-based descriptions. In deriving an overall standard from component standards, there was evidence that a linear compensatory model accurately captures the judges' aggregation strategies.

[1] proficiency in performing tasks that are central to the MOS. The tasks represent the core of the job and are the primary definers of the MOS.

[2] refers to the degree to which a synthetic equation was able to predict performance in the specific job for which it was developed

[3] refers to the degree to which performance in a job is better predicted by the synthetic equation developed for that job than by the synthetic equations developed for other jobs

TABLE 1.1
RELIABILITY ESTIMATES OF PHASE I JOB DESCRIPTION RATINGS AND VALIDITY RATINGS

		MOS	
	11B	63B	71L
Task Category Importance for			
Core Technical Proficiency	.52	.36	.40
Overall Job Performance	.52	.43	.44
Job Activity Importance			
Core Technical Proficiency	.36	.23	.43
Overall Job Performance	.36	.25	.34
Attribute			
Validity (Soldiers)	.31	.34	.45
Validity (Psychologists)	.42	.55	.52

TABLE 1.2
COMPARING SYNTHETIC AND EMPIRICAL COMPOSITES OBTAINED IN PHASE I

Composites	Mean Absolute Validity*	Discriminant Validity
Empirical Composites	.67	.17
Synthetic Composites		
Task Category	.55	.01
Job Activity	.53	.01
Attribute (Soldiers)	.52	.02
Attribute (Psychologists)	.58	.04

Note: * averaged across the three Phase I MOS

TABLE 1.3
METHODS OF JUDGING IMPLIED PERCENT OF SOLDIERS PERFORMING AT EACH LEVEL

				Percent		Percent	
MOS	Performance			Unacceptable		Outstanding	
	Dimension	Method	N	Mean	SD	Mean	SD
11B	General Soldiering	Soldier	80	8.0	5.3	12.4	9.6
		Task	81	21.0	14.9	7.7	9.4
		Incident	80	6.3	13.3	11.6	15.0
63B	General Soldiering	Soldier	49	8.4	6.9	16.3	18.6
		Task	50	23.0	14.6	11.0	12.1
	Basic Maintenance	Soldier	49	12.6	12.8	11.0	10.5
		Task	50	6.0	7.4	34.4	20.8
		Incident	49	4.4	16.3	8.8	12.6
71L	General Soldiering	Soldier	47	10.7	10.5	10.7	9.7
		Task	51	18.9	12.6	11.9	11.6
	Typing	Soldier	47	8.1	5.5	12.0	13.8
		Task	51	35.7	15.6	7.3	7.6
		Incident	52	10.8	14.7	9.2	12.2
	Other Clerical	Soldier	47	10.3	13.0	10.8	14.4
		Task	50	35.7	18.7	8.0	7.9
		Incident	52	4.6	12.4	4.8	5.6

Phase II Objectives and Research Questions: Job Description

At the conclusion of Phase I, we demonstrated that synthetic validation yielded valid predictions for all three jobs. One principal objective of Phase II was to compare the alternative job analytic methods on a number of distributional and psychometric properties that could serve as indicators of their comparative value for doing synthetic validation. Consequently, certain comparative questions were addressed using data collected in the Phase II workshops. These are the results that must be used to identify the method of choice for operational synthetic validation. Three major parameters characterize the alternative methods: type of descriptor, type of response scale, and type of expert judge. The relevant research questions are as follows.

(1) For each descriptor type, are there gaps in the "coverage" for specific MOS, as evidenced in the open-ended responses or the frequency of item endorsement?

(2) What are the comparative levels of inter-judge agreement by type of item, type of judge, type of response scale?

(3) Comparatively speaking, how well do the different instruments discriminate among MOS?

(4) What response scale, or scale composite, yields the highest reliability and greatest discrimination?

(5) Which judges yield the highest reliability and across-MOS discrimination?

(6) Are there any critical interactions between type of judge and type of descriptor relative to reliability or discriminability?

(7) Which method of synthetic validation produces the highest estimated validity for each MOS in the Phase II sample?

(8) For which method(s) do the synthetically estimated validities match the Project A empirical validities most closely?

(9) Which method yields the maximum differential prediction?

(10) Which method yields the level of differential prediction that most closely matches the Project A results?

We will return to these research questions in Chapter 7.

Phase II Objectives and Research Questions: Standard Setting

Also in Phase I, meaningful standards were obtained for the three jobs. In Phase II, we have attempted to refine the standard setting methods to yield better agreement among the judges and greater convergence across methods. For the three basic standard setting instruments the relevant research questions are:

(1) For each instrument, to what extent did different types of judges (NCO vs Officer, FORSCOM vs TRADOC) differ in terms of the mean levels of the standards that they set or the level of agreement (as measured by the standard deviation of the judgments across judges or by reliability estimates)?

(2) For the Critical Incident and the Task instruments, were the post-Delphi judgments significantly different from the initial judgments in terms of means and agreement levels?

(3) For the Task instrument, were there differences (in mean levels, agreement levels, and Delphi changes) among standards based on the hypothetical soldier ratings, the detailed information percent-go ratings, and the abbreviated information percent-go ratings?

(4) Were there differences among the different instruments (and the three different approaches within the task-based instrument) in terms of means, agreement levels, Delphi effects, and discrepancies across judge types?

For the exercise on combining multiple standards, the basic questions for analysis were:

(5) To what extent did a compensatory model explain the judges ratings better than a multiple hurdles model?

(6) Did the judges give equal weight to each performance dimension?

(7) Were the overall ratings significantly higher or lower than the simple average?

We will return to these research questions in Chapter 9.

Summary of Report Contents

In Chapter 2 we describe the methods and procedures used in conducting the workshops for Phase II. Chapter 3 summarizes the open-ended written and verbal comments provided by the workshop participants. Chapter 4 presents analyses of the questionnaires designed to measure the participants' knowledge of and experience with the Phase II MOS. Chapter 5 contains the analyses of the job description questionnaires: the Task, Activity, Hybrid, and Attribute Validity approaches. Chapter 6 describes how the synthetically formed prediction equations were put together and evaluates the various types of equations in terms of validity for predicting job performance. Chapter 7 presents a brief summary of the evidence about the various synthetic validation models that is contained in the prior four chapters. Chapter 8 contains standard setting results and Chapter 9 summarizes these findings and contains conclusions and recommendations for Phase III.

Apart from the results described in this volume, additional material on Phase II is presented in two other companion volumes. Volume II (Peterson, Owens-Kurtz, Hoffman, Arabian, & Whetzel, In preparation) contains appendices that present additional detailed results. In Volume III (Wise, Peterson, Hoffman, & Arabian, In preparation), we include all forms and instruments that were used in Phase II.

CHAPTER 2: METHOD AND PROCEDURES

Janis S. Houston (Personnel Decisions Research Institute, Inc)

Phase II data collection workshops were conducted from mid-January through the end of March 1989 at 10 Army installations throughout the continental United States. These workshops were eight hours in duration and ranged in size from 6 to 18 participants, with an average group size of approximately 12. Separate workshops were held for NCOs and officers and, except for rare instances, separately by MOS.

Description of Sample

General Description of Sample

Seven MOS were under study in the Phase II workshops. Three of these were Project A "Batch A" MOS. They were:

- Armor Crewman (19E/K)
- Motor Transport Operator (88M)
- Medical Specialist (91A/B)

The other four MOS were "Batch Z" MOS in Project A. These four were:

- MANPADS Crewmember (16S)
- Utility Helicopter Repairer (67N)
- Unit Supply Specialist (76Y)
- Food Service Specialist (94B)

There were two basic types of participants requested for the workshops. The first were NCOs and officers assigned to the Directorate of Training and Doctrine (DOTD) and other personnel who help define doctrine and prepare training plans for each of the seven MOS. The second were FORSCOM NCOs in each of the seven MOS and the officers who supervise the first-term soldiers in the MOS.

There were six DOTD data collection sites, all TRADOC posts, including:

- Fort Bliss (16S)
- Fort Eustis (67N, 88M)
- Fort Knox (19E/K)
- Fort Lee (76Y, 94B)
- Fort Rucker (67N)
- Fort Sam Houston (91A/B)

The four FORSCOM sites were:

- Fort Polk (16S, 67N, 91A/B)
- Fort Riley (16S, 19E/K, 67N)
- Fort Sill (76Y, 94B)
- Fort Stewart (16S, 88M)

Sample Sizes by MOS, Rank, and Command

A total of 476 personnel were requested for this data collection effort. Of this number, 408 (86%) participated in the workshops. Table 2.1 presents the total sample of participants, requested and obtained, by MOS and site. As can be seen in this table, all seven MOS are represented by both TRADOC and FORSCOM sites, and occasionally by more than one site within TRADOC or FORSCOM.

Table 2.2 shows the total number of participants obtained for each MOS by Rank and Command. For most MOS we had a total N between 50 and 60. Since there are relatively few 16S and 67N Army-wide, we somewhat over-tasked sites for personnel in these MOS, to ensure sufficient representation in our sample. In the case of FORSCOM 16S, we actually obtained slightly higher than the number requested.

TABLE 2.1.

NUMBER OF PHASE II PARTICIPANTS BY MOS AND SITE (REQUESTED NUMBER IN PARENTHESES)

Site	MOS	NCOs	Officers	Totals
Ft. Bliss	16S	(12) 11	(12) 11	(24) 22
Ft. Eustis	67N	(12) 9	(12) 10	(24) 19
	88M	(12) 13	(12) 10	(24) 23
Ft. Knox	19E/K	(12) 14	(12) 9	(24) 23
Ft. Lee	76Y	(12) 12	(12) 10	(24) 22
	94B	(12) 9	(12) 6	(24) 15
Ft. Rucker	67N	(12) 13	(0) 0	(12) 13
Ft. S. Houston	91A/B	(12) 15	(12) 11	(24) 26
TRADOC Subtotals		(96) 96	(84) 67	(180) 163
Ft. Polk	16S	(10) 11	(10) 10	(20) 21
	67N	(10) 6	(10) 11	(20) 17
	91A/B	(18) 15	(18) 18	(36) 33
Ft. Riley	16S	(18) 18	(18) 6	(36) 24
	19E/K	(18) 16	(18) 15	(36) 31
	67N	(10) 5	(10) 4	(20) 9
Ft. Sill	76Y	(18) 16	(18) 13	(36) 29
	94B	(18) 16	(18) 13	(36) 29
Ft. Stewart	16S	(10) 13	(10) 10	(20) 23
	88M	(18) 19	(18) 10	(36) 29
FORSCOM Subtotals		(148) 135	(148) 110	(296) 245
TOTALS		(244) 231	(232) 177	(476) 408

TABLE 2.2

NUMBER OF PHASE II PARTICIPANTS BY MOS

	NCOs	Officers	
MOS	TRADOC/FORSCOM	TRADOC/FORSCOM	Totals
16S	11/42	11/26	90
19E/K	14/16	9/15	54
67N	22/11	10/15	58
76Y	12/16	10/13	51
88M	13/19	10/10	52
91A/B	15/15	11/18	59
94B	9/16	6/13	44
Totals	96/135	67/110	408

Demographics of Sample

Tables 2.3 - 2.5 display the demographics of workshop participants, separately for each MOS. The demographic variables included in these tables are: rank, pay grade, time in service, and time in MOS.

Although somewhat redundant with prior tables, the "rank" variable is included here (Table 2.3) to demonstrate that we had civilian participants as well as NCOs and officers. This occurred only at TRADOC sites, where the DOTD and related functions are often performed by civilians as well as military personnel. When this was the case, we requested that the civilians be sent to whichever workshop (NCO versus officer workshop) they themselves, and our Point of Contact deemed appropriate. Thus, the NCO/officer numbers in Tables 2.1 and 2.2 include these few civilians; in Table 2.3, they are broken out.

Although we initially requested only soldiers in pay grade (Table 2.4) E7-E9 for NCO participants, and 02-04 for officers, we reduced this requirement when we learned how few NCOs and officers at these levels were available at some sites. We were assured by the on-site Point-of-Contact that all personnel in grades lower than requested (e.g., E5, E6) that were tasked to attend the workshops were very knowledgeable in the target MOS.

TABLE 2.3

DEMOGRAPHICS OF SAMPLE: RANK

MOS	Civilian	NCO	Officer	Total
16S	2	53	35	90
19E/K	9	26	19	54
67N	1	34	23	58
76Y	2	28	21	51
88M	4	32	15	52 (1 Unknown)
91A/B	0	30	29	59
94B	0	26	18	44
Total	18	229	160	408 (1 Unknown)

TABLE 2.4

DEMOGRAPHICS OF SAMPLE: PAY GRADE

MOS	E4-6	E7-9	W1-4	01-02	03-05	GS8-9	GS10-12
16S	45	8	0	20	15	1	1
19E/K	8	18	0	12	7	3	6
67N	22	12	10	6	7	1	0
76Y	16	12	2	10	9	0	2
88M	20	12	0	8	7	0	4
91A/B	10	20	2	13	14	0	0
94B	13	13	4	6	8	0	0
Total	134	95	18	75	67	5	13

TABLE 2.5

DEMOGRAPHICS OF SAMPLE: TIME IN SERVICE AND TIME IN MOS, IN YEARS

	Time in Service		Time in MOS	
MOS	Mean	SD	Mean	SD
16S	10.2	5.8	6.8	4.2
19E/K	11.4	7.4	9.7	7.5
67N	11.8	5.9	8.7	5.9
76Y	11.6	6.4	8.1	5.9
88M	11.4	6.4	10.1	6.8
91A/B	12.2	6.6	10.7	7.2
94B	12.9	5.5	10.2	7.0

Workshop Procedures

General Description of Procedures

At the beginning of each eight-hour workshop, participants were given an overview of the project and then briefed on the day's activities. A Privacy Act statement was distributed and read, and a Background Information sheet was completed by each participant. Participants were given an opportunity to ask questions about the project and the workshop. Two questionnaires were then administered to measure participants' experience and familiarity with their MOS. These questionnaires will be described in the following section.

The remainder of the workshop was divided into two sets of exercises: four job component questionnaires/exercises comprised one set and five standard setting exercises comprised the other set. All workshops presented the set of job component exercises first, followed by the set of standard setting exercises. The order of administration within set, however, varied across workshops.

Each of these exercise forms will be described in the next section, followed by an explanation of the variations in exercise order. Volume III of this report contains a sample copy of all workshop forms.

Description of Forms to Assess the Knowledge and Experience of Workshop Participants

There were two questionnaires designed to assess the knowledge and experience of the workshop participants. These were administered at the beginning of the workshop (after the introductory activities described earlier). A description of these questionnaires follows.

Job Familiarity Questionnaire. This questionnaire was designed to measure participants' knowledge of critical MOS tasks. For each MOS, a sample of around 15 test items were taken from the Project A task-based or school-based job knowledge test. These items were assembled as the Job Familiarity Questionnaire, which typically took approximately 10 minutes for workshop participants to finish.

Job History Questionnaire. This questionnaire was developed during Project A to assess how frequently and recently individuals had performed critical first-tour tasks for their MOS. The questionnaire includes 30 critical first-tour tasks identified by the Project A job analysis. Participants were asked to indicate, for each task, how frequently and how recently they had performed, supervised, taught, or graded the task. They typically took about 15 minutes to complete this questionnaire.

Description of Job Component Exercises

There were four job component exercises. The order of their administration varied across workshops. (See later section on administration order.) Each of these exercises is described below.

MOS Task Questionnaire. There were 96 task categories in this questionnaire. In the instructions, participants were instructed to consider soldiers with 18 months on-the-job experience in their MOS and in the full range of duty assignments as they rated the relative frequency of each task. After completing the Frequency ratings, participants were asked to make three Importance ratings for each task: 1) the task's importance for Core Technical Proficiency[1]; 2) its importance for General Soldiering[2]; and 3) its importance for overall job performance. Most participants took 30 to 45

[1] is made up of tasks that are central to the MOS. The tasks represent the core of the job and are the primary definers of the MOS.

[2] individuals in every MOS are responsible for being able to perform a variety of general soldiering tasks. These are referred to as "Common Tasks."

minutes to complete this questionnaire. A copy of the questionnaire is found in Volume III, Attachment 6.

MOS Activity Questionnaire. There were 53 activities in this questionnaire. Participants were given instructions that were identical to those for the MOS Task Questionnaire, i.e., there were four ratings to be made for each activity. The average length of time to complete this questionnaire for the various groups was 20 to 30 minutes. A copy of the questionnaire is found in Volume III, Attachment 5.

General Task and Activity ("Hybrid") Questionnaire. This instrument (see Volume III, Attachment 9) combined elements from the Task and the Activity questionnaires into 38 task/activity statements. Participants were asked to make five ratings for each job element (task/activity). The first four ratings were identical to those in the Task and the Activity questionnaires. The fifth rating was relative difficulty of learning and performing each job element satisfactorily. This questionnaire typically took 20 to 25 minutes to complete.

Attribute Validity Ratings and Rankings. At the beginning of these exercises, the workshop leader explained the terms "attribute," "job performance," and "validity" to the participants. A booklet containing definitions of 30 attributes was provided. Graphs were used to illustrate high and low validity. Participants were then asked to rate the validity of each attribute for up to five different job performance areas:

- Core Technical Proficiency - represents the proficiency with which the soldier performs the tasks that are "central" to the MOS, i.e., the specific job that the soldier performs. The tasks represent the core of the job and are the primary definers of the MOS.

- General Soldiering Proficiency - represents the proficiency with which the soldier performs a variety of general soldiering tasks. For example, determines grid coordinates on military maps and determines a magnetic azimuth using a compass.

- Effort and Leadership - reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers.

- Personal Discipline - reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates integrity in day-to-day behavior, and does not create disciplinary problems. People who rank high on this area show a commitment to high standards of personal conduct.
- Physical Fitness and Military Bearing - represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

After completing this, participants filled out an Attribute Ranking Questionnaire, wherein they rank-ordered the attributes from high to low validity for predicting overall job performance. Combined, the Attribute Validity Ratings and Rankings took approximately one hour to complete, including directions.

Description of Standard Setting Exercises

There were five standard setting exercises. Again, the order of their administration was varied across workshops (discussed in the next section), but the entire set of these exercises always came after the job component exercises.

Prior to beginning the standard setting exercises, the workshop leader presented definitions of four levels of performance:

- Unacceptable - Soldiers who consistently perform like this do not belong in the Army. Their performance is hurting the Army, and it does not seem likely that additional training could bring their performance up to acceptable levels. Such soldiers should be discharged early.
- Marginal - Soldiers who consistently perform like this need remedial training. Their current performance is of little or no benefit to the Army. Unless they receive additional training and improve their performance, they should be barred from re-enlistment.
- Acceptable - Soldiers who consistently perform like this are doing an adequate job. They are making positive contributions to the Army. They should be allowed to re-enlist.
- Outstanding - Soldiers who consistently perform like this are doing extremely well. They are making exceptional contributions to the Army and are good examples to other soldiers. They should be given special incentives to encouraged them to re-enlist and should be given consideration for early promotion.

These definitions were used for all five exercises. The form used in the field test is shown in Volume III, Attachment 10. The form always accompanied each of the five exercises. A brief description of these exercises follows.

Soldier-Based Exercise. For the Soldier-Based Exercise (see Volume III, Attachement 11), participants were presented with a list of job performance categories that had been identified for their MOS. For each category, they were asked to indicate the percent of incumbents they would rate as unacceptable, marginal, acceptable, or outstanding. No normative data was provided to guide these ratings; instead, judges were asked to rely on their own experience. This exercise usually took 10 to 15 minutes to complete.

Critical Incident-Based Exercise. For this exercise, participants were provided with examples of Army-wide and MOS-specific critical incidents collected for Project A. The incidents had been sampled carefully to ensure coverage of all key performance factors and levels of performance. The incidents were presented in order from least effective to most effective (based on retranslation ratings) within each performance factor. Participants were asked to rate each incident as indicative of unacceptable, marginal, acceptable, or outstanding performance. It typically took 25 to 35 minutes to complete this exercise. (See Volume III, Attachement 12 for a copy of this exercise).

Task-Based Exercises: Detailed and Abbreviated Forms. In the Task-Based Exercises (see Volume III, Attachment 13), participants received a questionnaire that included detailed performance information, including scoresheets, on some tasks (Detailed Form) and no performance information on other tasks (Abbreviated Form). All participants completed both forms, except for 94B SMEs, for whom we did not have Detailed performance information available from Project A.

On the Detailed Form, participants were asked to indicate whether a hypothetical soldier's performance on each task and overall was unacceptable, marginal, acceptable, or outstanding. In addition, each participant also indicated, for each task, the minimum percent of steps that should be passed for marginal, acceptable, and outstanding performance. On the Abbreviated Form, participants were asked only to indicate, for each task, the minimum percent of steps that should be passed for marginal, acceptable, and outstanding performance.

The Abbreviated Form took 15 to 20 minutes to complete; the Detailed Form ranged from 30 to 50 minutes.

Overall Standard Setting Questionnaire. This questionnaire was always given last. Participants were given hypothetical soldiers' standing (unacceptable, marginal, acceptable, or outstanding) on several performance dimensions, and were asked to make an overall rating of Core Technical Proficiency. This questionnaire typically took 15 to 25 minutes to complete. (See Volume III, Attachment 14 for copy of this questionnaire).

Order Variations Across Workshops

The order of administration was varied across workshops for both the set of job component exercises and the set of standard setting exercises. Table 2.6 presents the order in which the job component exercises were administered for each type of workshop, where type of workshop is defined by MOS, Rank, and Command. (Separate workshops were conducted for each MOS/Rank/Command group). Table 2.7 displays the administration order for the standard setting exercises. As can be seen in these two tables, an attempt was made to balance order within set across workshops.

TABLE 2.6

ADMINISTRATION ORDER FOR JOB COMPONENT EXERCISES

	TRADOC		FORSCOM	
MOS	NCOs	Officers	NCOs	Officers
16S	1*	2	4	3
19E/K	2	1	3	4
67N	1	4	2	3
76Y	2	3	1	4
88M	3	4	2	1
91A/B	3	2	4	1
94B	4	3	1	2

*Order of Questionnaires/Exercises:

1 = Task, Activity, Attributes, General

2 = General, Attributes, Task, Activity

3 = Activity, Task, Attributes, General

4 = General, Attributes, Activity, Task

TABLE 2.7

ADMINISTRATION ORDER FOR STANDARD SETTING EXERCISES

	TRADOC		FORSCOM	
MOS	NCOs	Officers	NCOs	Officers
16S	S,T(A,D),C*	S,C,T(D,A)	C,T(D,A),S T(D,A),C,S	C,S,T(A,D) T(A,D),S,C
19E/K	C,T(D,A),S	T(A,D),S,C	S,T(D,A),C	S,C,T(D,A)
67N	T(A,D),S,C	S,C,T(D,A)	C,T(D,A),S S,C,T(D,A)	D,T(A,D),S T(A,D),S,C
76Y	S,T(D,A),C	T(A,D),C,S	S,T(A,D),C	C,T(D,A),S
88M	C,T(A,D),S	S,C,T(D,A)	S,C,T(D,A)	C,T(A,D),S
91A/B	T(A,D),S,C	S,T(D,A),C	C,T(D,A),S	T(A,D),C,S
94B	T(D,A),C,S	T(A,D),S,C	S,T(A,D),C	S,C,T(D,A)

*Notes:

S = Soldier-Based Exercise

T = Task-Based Exercise (A=Abbreviated D=Detailed)

C = Critical Incident-Based Exercises

Also note that two orders are given for 16S and 76N, FORSCOM. This was to accommodate there being two FORSCOM sites for these MOS. Thus, a different order was prescribed for every workshop.

Standard Setting Delphi Sessions

In each workshop, a delphi session was conducted for either the Task-Based or Critical Incident-Based Exercise. Each delphi session was preceded by a short break, allowing the workshop leader to collect the initial rating sheets and tally the responses for two of the performance dimensions.

For the Task-Based delphi sessions, the workshop leader identified the four or five hypothetical soldiers for which there was the greatest disagreement on each of the two dimensions. (Where disagreement is defined as the number of judges giving other than the modal rating.) For the Critical Incident-Based delphi sessions, the workshop leader identified the four or five incidents within each of the first two dimensions for which there was the greatest disagreement.

The workshop leader then presented the results for each of the discrepant soldiers or incidents. Participants were asked to state specific negative or positive consequences of the indicated performance to support their particular rating or to otherwise explain their rating strategy. If suggestions were not forthcoming, the workshop leader suggested some general types of outcomes, including: harm to the soldiers or others, damage to equipment, likelihood of mission failure, contribution to mission success, contribution to the successful performance of other soldiers.

Following the discussion, participants were asked to complete a second copy of the Task-Based or Critical Incident-Based instrument. During the discussion and the readministration, the workshop leader wrote down the key rationales provided by the participants. The rationales will be compiled and examined for possible use in developing rater training for Phase III workshops.

Table 2.8 presents the delphi session design, i.e., which delphi was conducted for each workshop.

TABLE 2.8

DELPHI SESSION DESIGN

MOS	TRADOC		FORSCOM	
	NCOs	Officers	NCOs	Officers
16S	T*	C	T,C	C,T
19E/K	C	T	T	C
67N	T	C	C,T	T,C
76Y	C	T	T	C
88M	T	C	C	T
91A/B	C	T	T	C
94B	C	T	T	C

*Notes: C = Delphi was conducted for Critical Incident-Based Exercise

T = Delphi was conducted for Task-Based Exercise

Where two letters are listed, there were two workshops for that group.

CHAPTER 3: SUMMARY OF PHASE II WORKSHOP REPORTS

John P. Campbell (Human Resource Research Organization)

This chapter presents a summary of the open-ended written comments and the verbal comments offered by the NCO and officer participants in the Phase II workshops. Recall that there were 36 workshops distributed over 10 posts and that there were 18 NCO workshops and 18 officer workshops. The 36 groups varied in size from 6-18 with most being in the 8-14 range.

Each of the job description and standard setting instruments provided space for open-ended comments about item content and suggestions for improving the procedure. The written comments from each workshop were summarized in a standard format by the project staff. The relevant verbal comments made during the workshop discussions were summarized in the trip reports filed by the respective workshop leaders, to the extent that their notes and recollections permitted.

This information was used to address the following general questions.

(1) What were the strengths and limitations of the workshop formats as viewed by the officer and NCO participants and by the workshop leaders (i.e., project staff)? Are revisions thought to be needed?

(2) What seemed to be the overall evaluations of each instrument by the participants?

(3) Are there clear opinions about specific revisions in the content of the various instruments?

(4) Are there clear opinions about specific revisions in the judgment or scaling procedures that were used?

The data were "analyzed" simply by having one project staff member who had not been at the workshops content analyze both the workshop reports and the trip reports. The results of this effort were reviewed for accuracy and reasonableness by three project staff who had been workshop leaders.

Results

Only the most frequent and the most relevant comments are summarized here. That is, comments that were made by only one person or that were not plausible grounds for revisions in workshop format, instrument content, or scaling techniques are not reported. Understandably, most of the written and verbal comments pertained to suggestions for revisions. When these are summarized in one place, the result is a tone which may sound more critical than it should. Consequently, these results should be interpreted in the context of the overall judgment by participants and workshop leaders that the workshops went very well and most participants found it an interesting activity.

Workshop Format and Procedure

In general, everyone seemed to think that the workshop procedures worked fairly well; however, we were probably guilty of trying to pack too much into one day. Interest sometimes lagged toward the end of the day and a consensus seemed to be reached "too quickly" in some of the Delphi sessions. The specific participant reactions that can be noted are the following.

- The time squeeze probably attenuated the effects of the Delphi discussions in some of the standard setting sessions. There may not have been enough time for the participants to discuss their reasons for assigning particular scale values.

- A number of participants did not understand or would not accept the purpose of the job description instruments. That is, they complained that many items did not pertain to their MOS, for which there were too few items. The explanation that the questionnaire was meant to be used for all MOS and, by design, only a subset of items would be relevant for a specific MOS was either not comprehended or not accepted. It may be the case that people feel uncomfortable in such a situation when only a few items can be used to describe an entire MOS that is their own.

- Some participants from the non-combat specialties felt that all the instruments placed too much emphasis on the combat MOS.

- A number of people objected to the "18 month first termer" as a prototype since such individuals are still too inexperienced to be performing the full range oĩ tasks in their MOS. Perhaps 24-30 months would be more appropriate.

Job Description Item Content

In general, no one complained about the length of the instruments. If anything, the problem was the opposite. There were a number of complaints that the task and activity items were too general and that they should be more specific. The additional points raised by the participants tend to fall into two categories -- suggestions that are not instrument specific and suggestions for specific item content for each instrument.

General points.

- Most participants tended to like the attributes more than tasks or activities because the former were more completely defined. The task and activity items were apparently defined so succinctly for some that they weren't always sure whether an item applied to their MOS or not, or how critical or difficult it would be.

- A number of respondents questioned whether the task and activities' item content reflected job content that was too advanced for the 18-month job incumbent.

- Many people complained that the task and activity questionnaires did not adequately cover their own MOS.

- Most of the suggestions for new attributes were in the form of traits such as integrity, courage, and loyalty, or value dimensions such as patriotism and faith. An exception was adaptability/flexibility.

- The majority of suggestions for new task or activity items were of a "common task" nature; however, they tended to be common tasks which were particularly relevant for the MOS of the individual making the suggestion.

- The next largest proportion of new items were really subtasks or subactivities of existing items, which goes back to the general desire for more specificity.

- There was some disagreement whether there should be items pertaining to supervision or coaching. The judges from 91A were definitely in favor of it. Most others were against.

Specific item suggestions. The following list is meant to avoid common tasks and items that are more specific versions of existing items.

<u>Task items</u>

- More aircraft related items.
- More items pertaining to the operation of communication equipment.
- Items pertaining to sanitation and hygiene.

<u>Activity items</u>

- Reading documents, manuals, messages.
- Transferring information from one person or group to another.
- Change medical problem solving to physical/mental health problem solving.
- Monitor and inspect.
- Facilitating social interactions.
- Matching names to objects.
- Recalling a sequence of events or steps.
- Standing for a long time.
- Making decisions.
- Working autonomously or without direct supervision.

Standard Setting

In general, and as expected, the standard setting exercise proved much more difficult than the job analysis exercises. The major issues and questions reflected in the open-ended reports seemed to be the following.

- In the task-based method a number of participants reacted negatively to setting standards on tasks that were not specifically in their MOS or with which they were not familiar. as when they were asked to make the correspondence between the displayed task and a similar task in their own MOS. This was the major issue.

- Some people had trouble with the Unsatisfactory, Marginal, Acceptable, and Outstanding (U, M, A, O) scale definitions. They were hard to use in the critical incident method because some participants did not want to make such judgments on the basis of one critical incident but had trouble adopting a frame of reference in which an incident represented a typical episode. Someone also suggested using the term "behavioral examples" rather than critical incidents as a label because the latter has its own meaning in the Army.

- A few people wanted to restate the definitions of U, M, A, and O in terms of the actions the Army would take with such job incumbents. A few others questioned the use of "Outstanding" as a critical category on the groups. The current situation doesn't really permit the demonstration of outstanding performance even if that is the individual's true score.

- There was some suggestion that if a critical incident or task-based method were used then the criticality ju1gment should be separated from the standard setting judgment. That is, the standard expected would be partly a function of the importance of the task (e.g., "acceptable" performance is at a higher level for more critical tasks).

- A number of people did not like the "percent" judgments in the task-based method. In general, the soldier-based method presented an easier scaling task.

Overall Comparisons

In general, the following conclusions seemed to be the majority view, if not the consensus.

- For the job analysis methods attributes, tasks, and activities were generally preferred in that order, but not unanimously. Because of their more complete definitions the attribute items were the least ambiguous to use. Task items were generally preferred to activity items because they were in MOS terms and were more specific. However, activity items were sometimes preferred because a higher proportion of them could be used for each MOS.

- Ratings were preferred to rankings for the attribute judgments. Some complained that the ranking method required them to remember too much.

- Very few comparative comments were made relative to alternative methods of standard setting, although the task-based methods received considerable negative comments, and the abbreviated task method received the most negative comments of all.

Possible Action Steps

Based on the reports discussed above, the following actions are suggested. These suggestions are independent of the actual empirical results obtained from the various instruments.

(1) The hybrid questionnaire should be dropped and either:

(a) the task and activities items should be summed, or

(b) the task questionnaire should be expanded.

(2) A significant number of common task items should be included to allow judges to give a more complete description of their MOS.

(3) A careful review of all specific item suggestions should be conducted for the purpose of one final attempt to add item content. Careful consideration should be given to replace general items with somewhat more specific ones. There could probably be at least 150-175 items without creating any difficulties, particularly in an operational setting when this would be the only instrument being used.

(4) The tenure specifications for the prototypic first term incumbent should be increased to 24 months.

(5) Serious consideration should be given to revising the performance level definitions in the standard setting exercises, particularly the use of unacceptable and outstanding.

(6) The comparison of the soldier-based and task-based methods should be made more efficient (e.g., with a single MOS) and more time should be allowed for discussion among the judges earlier in the day. Asking judges to use the task-based method to estimate standards outside their own MOS probably won't work.

CHAPTER 4: MEASURES OF THE KNOWLEDGE AND EXPERIENCE OF WORKSHOP PARTICIPANTS: JOB FAMILIARITY AND JOB HISTORY QUESTIONNAIRES

Cynthia K. Owens-Kurtz (PDRII) and Carolyn Hill-Fotouhi (HumRRO)

In Phase I, analyses were completed which assessed the impact of several characteristics on the "fidelity" of participants' responses to the Attribute Validity Task, and Activity Rating Questionnaires. Fidelity refers to the agreement between a participant's response profile and the mean response profile based on all other participants. Three characteristics were investigated: job experience, job knowledge, and general aptitude. The measure of job experience included job tenure and task experience, assessed with the Job History Questionnaire. Job knowledge was measured with the Job Familiarity Questionnaire and the Army Skill Qualification Test (SQT) score. General aptitude consisted of the Armed Forces Qualification Test (AFQT) composite score and educational attainment.

Results reported by Szenas and McHenry (1989) suggested that job knowledge and general aptitude are related positively to the fidelity of a participant's responses to the Task and Activity Questionnaires. The possibility of improving the quality of job description by screening participants on these variables was proposed, and a recommendation was made for further research.

In Phase II, the Job Familiarity and Job History Questionnaires were investigated further to determine the feasibility of using scores on these instruments to screen workshop participants. The questionnaire content and scoring and the results of our analyses are described below.

Description of Instruments

Two questionnaires were used to assess knowledge and experience in the MOS being studied. The Job Familiarity Questionnaire and the Job History Questionnaire are described separately.

Job Familiarity Questionnaire

This questionnaire was designed to measure participants' knowledge of critical MOS-specific and Army-wide tasks. For each MOS, a sample of 15 multiple-choice test items were taken from the Project A task-based or school-based job knowledge test. The content of the questionnaire was approximately balanced between MOS-specific and Army-wide items. (See Volume III, Attachments 5 for a copy of the Job Familiarity Questionnaire for each MOS.)

Job History Questionnaire

The Job History Questionnaire required participants to indicate how recently and frequently they had performed, supervised, or educated others on 28 to 30 MOS-specific and Army-wide job tasks. Items on the questionnaire were brief descriptions of first-tour tasks identified as critical during the Project A job analysis. (See Volume III, Attachment 4 for copy of questionnaire.) For each item, participants were asked if they had: (1) performed this task myself [Self], (2) directly supervised others doing this task [Supervised], (3) taught others to do this task [Taught], (4) tested or graded others on this task [Tested], and (5) wrote/revised manuals or tests on this task [Wrote]. They used the following 6-point scale to respond to the 5 queries about each item:

0 = Never
1 = 1 or more times, but more than one year ago
2 = 1-4 times in the last year
3 = 5-10 times in the last year

4 = 11-20 times in the last year

5 = More than 20 times in the last year.

Editing and Handling of Missing Data

Job Familiarity Questionnaire

Because the Job Familiarity Questionnaire is a multiple-choice knowledge test, a missing value is equivalent to an incorrect answer. Thus, all missing values were replaced with zeros for all analyses. Of the data, none of the 398 participants was missing more than four responses.

Job History Questionnaire

The number of items on the Job History Questionnaire varied by MOS from 140 (28 tasks x 5 experience categories) to 150 (30 tasks x 5 experience categories). A total of 405 individuals provided this data. In order to have a consistent standard, data were screened for cases with greater than 14 missing item responses, or 10% of the shortest questionnaire. Only one case, a Food Service Specialist Officer, exceeded this standard and was dropped, so the final sample size was 404. All remaining missing values were replaced with zeros for all analyses.

Results

Job Familiarity

Keys to the Job Familiarity Questionnaire for each MOS were developed and applied to the item responses for each soldier. Correct responses were scored 1; incorrect responses were scored 0. Each participant's score was the sum of the item scores, or the number of correct responses. A maximum of 15 points were possible. Obtained scores ranged from 4 to 15. Table 4.1 presents descriptive statistics and reliabilities for each MOS. Both alpha and odd-even reliabilities are presented for comparison. The tests included both MOS-specific and Army-wide items, and could not be considered strictly homogeneous. For four MOS, the internal consistency estimates of reliability (Alpha) were quite reasonable (> .4) for a test of this length. The reliability estimate for Motor Transport Operators (88M), however, are rather low, and the estimates are essentially zero for Helicopter Repairers (67N) and Unit Supply Specialists (76Y). The negative estimates for 76Y resulted from negative correlations between scores on several items (56 of 105 correlations were negative). This reflects either extreme heterogeneity in job content or changes in policy and procedures that affected the "correctness" of alternative responses. In order to identify possible homogeneous subsets of items, we performed a cluster analysis of the 15 76Y test items using the Ward method of linkage with 1-correlation coefficient as the distance measure (Wilkinson, 1988). Two clusters of items, with 8 and 7 items each, were identified and the reliabilities were recomputed for these subsets. The Spearman-Brown corrected split-half correlations were .61 and .36, respectively. These results seem to support the prior supposition about heterogeneity of the items. We reviewed item content and could discern no pattern in the way the items split into the two subsets. In any event, based on these low reliabilities and on the inconsistent pattern of reliabilities across MOS, we decided it was not feasible to use Job Familiarity scores as an indicator of job knowledge.

Job History

As described above, the Job History Questionnaire for each MOS consisted of 28 to 30 tasks. For each task, participants indicated their level of experience in five categories. Thus, the total number of items varied from 140 to 150 across the MOS. Subtest scores were computed for each of the five experience categories by summing the responses to each task for the category. Total scores were the sum of the five subtest scores. Experience level was rated on a 0 to 5 scale, thus the minimum possible total score for each MOS was 0, and the maximum possible total score ranged from 700 to 750. Obtained scores ranged from 0 to 552 across MOS.

TABLE 4.1
DESCRIPTIVE STATISTICS AND RELIABILITY FOR THE
JOB FAMILIARITY QUESTIONNAIRE, PHASE 2

MOS	N	MEAN	SD	ALPHA	ODD-EVEN[1]
16S	90	11.4	2.3	.558	.443
19K	54	12.3	2.0	.566	.638
67N	57	9.1	1.7	.071	.043
76Y	49	10.3	1.4	-.247	.000
88M	51	9.9	1.7	.238	.268
91A	59	9.7	2.1	.437	.361
94B	38	10.9	2.1	.467	.369

Note. The Job Familiarity Questionnaire for each MOS has 15 items.

[1] Spearman-Brown formula applied.

Table 4.2 presents the mean intercorrela*'*ons of the subtest and total scores across MOS. The subtest Wrote has low and sometim.. negative intercorrelations with the other four subtests and the total score. The subtest and total score intercorrelation matrices for each MOS appear in Volume II, Appendix A.

Table 4.3 contains the alpha and odd-even reliabilities for the Job History total scores for each MOS. The alpha reliabilities are moderately high, ranging from .764 to .895. Table 4.3 also presents descriptive statistics for FORSCOM, TRADOC, and combined samples for total Job History scores. Mean total scores range from 108 (MANPADS Crewmembers, TRADOC) to 314 (Armor Crewmembers, FORSCOM).

Subtest descriptive statistics and reliabilities appear in Volume II, Appendix B. With the exceptions of Helicopter Repairers and Medical Specialists, FORSCOM participants report higher mean levels of experience than TRADOC participants in the categories Self and Supervised. With the exception of MANPADS Crewmembers, TRADOC participants report higher mean levels of experience than FORSCOM participants in the category Wrote. No consistent pattern emerges for the categories Taught and Tested.

Judge subgroups were formed based on the Job History total scores. In the interest of forming two groups of about equal size, we used the median total score as a cut, with participants who scored at the median included in the high group. Table 4.4 presents cross-tabulations of the number of participants above and below the cut score by command. No consistent FORSCOM or TRADOC advantage emerges across MOS. However, within MOS there are clear discrepancies in the distributions by command. For two MOS, FORSCOM participants are cverrepresented in the "above" cut score group and TRADOC participants are overrepresented in the "below" cut score group (MANPADS Crewmembers: chi squared = 30.441, p < .001; Armor Crewmembers: chi squared = 21.888, p < .001). For Medical Specialists the reverse is true, with TRADOC overrepresented in the "above" group and FORSCOM overrepresented in the "below" group (chi squared = 3.931, p < .05). The remaining MOS have relatively even splits.

TABLE 4.2
MEAN INTERCORRELATIONS OF JOB HISTORY SUBTEST
AND TOTAL SCORES ACROSS MOS

	SELF	SUPER-VISED	TAUGHT	TESTED	WROTE
SELF	-----				
SUPERVISED	0.825	-----			
TAUGHT	0.787	0.830	-----		
TESTED	0.638	0.694	0.850	-----	
WROTE	0.041	0.052	0.152	0.226	-----
TOTAL	0.865	0.904	0.949	0.884	0.256

TABLE 4.3
DESCRIPTIVE STATISTICS AND RELIABILITIES FOR JOB HISTORY TOTAL SCORES BY COMMAND AND MOS

		COMMAND			
MOS		FORSCOM	TRADOC	OVERALL	NUMBER OF ITEMS
16S	N	68	22	90	150
	X	269.985	107.727	230.322	
	SD	107.529	39.277	118.260	
	Alpha			.895	
	Odd-Even[1]			.947	
19K	N	31	23	54	145
	X	314.484	111.000	227.815	
	SD	117.612	66.459	141.345	
	Alpha			.863	
	Odd-Even			.960	
67N	N	25	31	56	140
	X	142.360	158.871	151.500	
	SD	70.420	101.876	88.846	
	Alpha			.886	
	Odd-Even			.952	
76Y	N	29	21	50	145
	X	204.586	194.000	200.140	
	SD	86.697	118.682	100.360	
	Alpha			.852	
	Odd-Even			.930	
88M	N	29	23	52	145
	X	211.241	177.174	196.173	
	SD	84.125	106.830	95.396	
	Alpha			.869	
	Odd-Even			.939	
91A	N	33	26	59	145
	X	211.061	286.923	244.492	
	SD	104.388	133.849	123.196	
	Alpha			.893	
	Odd-Even			.954	
94B	N	28	15	43	140
	X	236.893	194.067	221.953	
	SD	75.856	89.388	82.395	
	Alpha			.764	
	Odd-Even			.901	

[1] Spearman-Brown formula applied.

TABLE 4.4

NUMBER OF JUDGES ABOVE & BELOW JOB HISTORY CUT SCORE BY COMMAND

16S: JOB HISTORY GROUP

	BELOW	ABOVE	TOTAL
FORSCOM			
N	22	46	68
Row %	32.4	67.6	
Column %	50.0	100.0	
TRADOC			
N	22	0	22
Row %	100.0	0	
Column %	50.0	0	
TOTAL	44	46	90

Note: chi squared = 30.441, p < .001

19K: JOB HISTORY GROUP

	BELOW	ABOVE	TOTAL
FORSCOM			
N	7	24	31
Row %	22.6	77.4	
Column %	25.9	88.9	
TRADOC			
N	20	3	23
Row %	87.0	13.0	
Column %	74.1	11.1	
TOTAL	27	27	54

Note: chi squared = 21.888, p < .001

87N: JOB HISTORY GROUP

	BELOW	ABOVE	TOTAL
FORSCOM			
N	11	14	25
Row %	44.0	56.0	
Column %	40.7	48.3	
TRADOC			
N	16	15	31
Row %	51.6	48.4	
Column %	59.3	51.7	
TOTAL	27	29	56

Note: chi squared = .321, p = .571

76Y: JOB HISTORY GROUP

	BELOW	ABOVE	TOTAL
FORSCOM			
N	12	17	29
Row %	41.4	58.6	
Column %	48.0	68.0	
TRADOC			
N	13	8	21
Row %	61.9	38.1	
Column %	52.0	32.0	
TOTAL	25	25	50

Note: chi squared = 2.053, p = .152

TABLE 4.4 CONTINUED

NUMBER OF JUDGES ABOVE & BELOW JOB HISTORY CUT SCORE BY COMMAND

88M: JOB HISTORY GROUP

	BELOW	ABOVE	TOTAL
FORSCOM			
N	13	16	29
Row %	44.8	55.2	
Column %	50.0	61.5	
TRADOC			
N	13	10	23
Row %	56.5	43.5	
Column %	50.0	38.5	
TOTAL	26	26	52

Note: chi squared = .702, p = .402

91A: JOB HISTORY GROUP

	BELOW	ABOVE	TOTAL
FORSCOM			
N	20	13	33
Row %	60.6	39.4	
Column %	69.0	43.3	
TRADOC			
N	9	17	26
Row %	34.6	65.4	
Column %	31.0	56.7	
TOTAL	29	30	59

Note: chi squared = 3.391, p = .047

94B: JOB HISTORY GROUP

	BELOW	ABOVE	TOTAL
FORSCOM			
N	12	16	28
Row %	42.9	57.1	
Column %	57.1	72.7	
TRADOC			
N	9	6	15
Row %	60.0	40.0	
Column %	42.9	27.3	
TOTAL	21	22	43

Note: chi squared = 1.149, p = .284

Given these results, the Job History score seems difficult to defend as a screening device for judge qualifications. In some MOS, nearly all TRADOC participants would be screened out, while in other MOS many FORSCOM participants would be screened out. In the remaining MOS, approximately equal numbers of TRADOC and FORSCOM participants would be screened out. If we are not willing to accept the finding that there are no qualified TRADOC judges in some MOS and few qualified FORSCOM judges in others, then Job History scores should probably not be recommended to screen participants as judges. In addition, the data presented in Chapter 5 indicate that TRADOC and FORSCOM participants differ very little with regard to reliability and convergent/discriminant validity analyses. Those findings do not track well with the Job History results.

Summary and Conclusions

Unfortunately, both instruments designed to quantify participant knowledge and experience have flaws. The Job Familiarity Questionnaire obtained very low reliability coefficients, and the Job History Questionnaire, while reliable, shows an inconsistent and illogical pattern of total scores for command groups across MOS. These initial findings suggested that further analyses such as those in Phase I on participant characteristics (Szenas & McHenry, 1989), would unduly capitalize on the low reliabilities and error. Thus, our Phase II analyses took a more detailed look at just two of the measures utilized in the Phase I research. It does seem unusual, however, that such large correlations were found between job knowledge and rating fidelity when one of the job knowledge measures, the Job Familiarity Questionnaire, has such low reliabilities. The other job knowledge measure, used in Phase I, SQT score, may have driven the relationship. Different MOS were involved, which also may account for the discrepancy. Whatever the explanation, the Phase II results do not support the use of Job Familiarity Questionnaire scores as a screen for participant job knowledge.

CHAPTER 5: ANALYSIS OF JOB DESCRIPTION QUESTIONNAIRES

Cynthia K. Owens-Kurtz (PDRII), Carolyn Hill-Fotouhi, R. Gene Hoffman (HumRRO), and Norman G. Peterson (PDRII)

This Chapter of the report is concerned with the evaluation of the four job description approaches: the Task, Activity, "Hybrid," and Attribute Validity Questionnaires. The Task, Activity, and Hybrid Questionnaires are similar in format and require very similar judgments from workshop participants. Essentially, SMEs are asked to describe their jobs in terms of the tasks and/or activities performed. In contrast, the Attribute Validity Questionnaires do not focus on job content. Instead, SMEs are asked to estimate the validity of attributes for performance in up to five broad job areas. In light of the similarities between the Task, Activity, and Hybrid Questionnaires, we will treat them together, separately from the Attribute Validity Rating and Ranking Questionnaires.

The research questions addressed in this chapter concern coverage of the MOS performance domain, the reliabilities of the various questionnaires, differences between rater types (i.e., Command and Rank), and discrimination between MOS. Data from Phase I were included in analyses of the Task, Activity, and Attribute Validity Rating and Ranking Questionnaires when applicable.

Description of Materials and Judgments

Task, Activity, and Hybrid Questionnaires

The Task Questionnaire consists of 96 task categories that describe job content in terms of the tasks performed. At the most general level, the tasks encompass four categories: (a) maintenance, (b) general operations, (c) administrative, and (d) combat. The task categories taxonomy is shown in Figure 5.1. The development of the Task

Questionnaire is described in detail in chapter 3 of the Phase I Synthetic Validation report (Chia, Hoffman, Campbell, Szenas, & Crafts, 1989).

The Activity Questionnaire is composed of general job behaviors that may be relevant for several specific job tasks. The questionnaire contains 53 items in the following general categories: (a) leadership, (b) communication, (c) information manipulation, (d) perceptual judgments, (e) problem solving, (f) operating equipment, (g) adjusting, (h) driving, (i) aiming, and (j) other physical actions. The job activity taxonomy is shown in Figure 5.2. A detailed description of the development of the Activity Questionnaire appears in chapter 3 of the Phase I report (Chia et al., 1989).

The Hybrid Questionnaire is a combination of the Task and Activity Questionnaires. The goal in developing a "hybrid" questionnaire was to combine the basic elements of the Task and Activity Questionnaires. Elements were derived from the similarity among task and activity items in terms of how each item is predicted by our array of attributes. The process began with the estimated correlations between task and activity job descriptors with the job attributes rated during the expert judgment phase of the project (see Chapter 6). Job descriptors were then treated as "variables" and attributes as "subjects" and intercorrelations among job descriptors were calculated. This intercorrelation matrix was then factored and the factor loadings were used to cluster the job descriptors. The result is a clustering of job descriptors based on the similarity of the relationships to predictor attributes. The clusters served as a starting point for defining hybrid elements and were redefined as necessary to straighten out apparent misrepresentations in the factor and cluster results. It contains 38 items which are more comprehensive, and therefore more abstract, than the Task and Activity Questionnaire items. The hybrid taxonomy is shown in Figure 5.3.

Figure 5.1 Task Category Taxonomy

I. Maintenance
A. Mechanical Systems Maintenance
1. Perform operator maintenance checks and services
2. Perform operator checks and services on weapons
3. Troubleshoot mechanical systems
4. Repair weapons
5. Repair mechanical systems
6. Troubleshoot weapons

B. Electrical and Electronic Systems Maintenance
1. Install electronic components
2. Inspect electrical systems
3. Inspect electronic systems
4. Repair electrical systems
5. Repair electronic components

II. General Operations
A. Pack and Load
1. Pack and load materials
2. Prepare parachutes
3. Prepare equipment and supplies for air drop

B. Vehicle and Equipment Operations
1. Operate power excavating equipment
2. Operate wheeled vehicles
3. Operate track vehicles
4. Operate boats
5. Operate lifting, loading, and grading equipment

C. Construct/Assemble
1. Paint
2. Install wire and cables
3. Repair plastic and fiberglass
4. Repair metal
5. Assemble steel structures
6. Install pipe assemblies
7. Construct wooden buildings and other structures
8. Construct masonry buildings and structures

D. Technical Procedures
1. Operate gas and electric powered equipment
2. Select, layout and clean medical/dental equipment and supplies
3. Use audiovisual equipment
4. Reproduce printed material
5. Operate electronic equipment
6. Operate radar
7. Operate computer hardware
8. Cook
9. Perform medical laboratory procedures
10. Conduct land surveys
11. Provide medical or dental treatment

Figure 5.1 Task Category Taxonomy (continued)

E. Make Technical Drawings
   1. Sketch maps, overlays, or range cards
   2. Produce technical drawings
   3. Draw maps and overlays
   4. Draw illustrations

III. Administrative

A. Clerical
   1. Type
   2. Prepare technical forms and documents
   3. Record, file, and dispatch information
   4. Receive, store, and issue supplies/equipment/other materials

B. Communication
   1. Use hand and arm signals
   2. Read technical manuals, field manuals, regulations, and other publications
   3. Use maps
   4. Send and receive radio messages
   5. Give oral reports
   6. Receive clients, patients, guests
   7. Give directions and instructions
   8. Write documents and correspondence
   9. Write and deliver presentations
   10. Interview
   11. Provide counseling and other interpersonal interventions

C. Analyze Information
   1. Decode data
   2. Analyze electronic signals
   3. Analyze weather conditions
   4. Order equipment and supplies
   5. Estimate time and cost of maintenance operations
   6. Plan placement or use of tactical equipment
   7. Translate foreign languages
   8. Analyze intelligence data

D. Applied Math and Data Processing
   1. Control money
   2. Determine firing data for indirect fire weapons
   3. Compute statistics or other mathematical calculations
   4. Provide programming and data processing support for computer operations

E. Control Air Traffic
   1. Control air traffic

Figure 5.1 Task Category Taxonomy (continued)

IV. Combat

A. Individual Combat
1. Use hand grenades
2. Protect against NBC hazards
3. Handle demolitions or mines
4. Engage in hand-to-hand combat
5. Fire individual weapons
6. Control individuals and crowds
7. Customs and laws of war
8. Navigate
9. Survive in the field
10. Move and react in the field

B. Crew-served Weapons
1. Load and unload field artillery or tank guns
2. Fire heavy direct fire weapons (e.g., tank main guns, TOW missile, IFV cannon)
3. Prepare heavy weapons for tactical use
4. Place and camouflage tactical equipment and materials in the field
5. Fire indirect fire weapons (e.g., field artillery)

C. Give First Aid
1. Give first aid

D. Identify Targets
1. Detect and identify targets

V. Supervision
A. Plan, operations
B. Direct/lead teams
C. Monitor/inspect
D. Lead
E. Act as a model
F. Counsel
G. Communicate
H. Train
I. Personnel administration

Figure 5.2. Job Activity Taxonomy

I. Leadership/Teamwork

A. Work in a team
B. Lead a team
C. Support/advise peers
D. Support/advise
E. Coach peers
F. Coach subordinates

II. Communication

A. Make oral reports (to individuals)
B. Make oral reports (to groups)
C. Relay oral instructions
D. Interview
E. Record information
F. Write brief messages
G. Write longer reports

III. Use Information

A. Monitor/interpret verbal messages
B. Recall verbal information
C. Monitor/interpret numerical information
D. Recall numerical information
E. Monitor/interpret figural information
F. Recall figural information
G. Follow oral directions
H. Follow written directions

IV. Perceptual Judgments

A. Judge size and distance
B. Judge location
C. Judge paths of moving objects

V. Problem Solving/Troubleshooting

A. Solve electrical system problems
B. Solve mechanical system problems
C. Solve logistical problems
D. Solve tactical maneuver problems
E. Solve administrative problems
F. Solve leadership problems
G. Solve medical problems
H. Solve communication problems

Figure 5.2. Job Activity Taxonomy (continued)

VI. Operate Equipment

A. Operate precision hand-held equipment
B. Operate hand-held tools
C. Operate hand-held power equipment
D. Operate large power equipment
E. Operate full keyboard
F. Operate numeric keyboard

VII. Adjust and Control

A. Adjust device using one limb
B. Adjust control device using multiple limbs

VIII. Drive

A. Drive tracked vehicle
B. Drive heavy wheeled vehicle
C. Drive light wheeled vehicle

IX. Aiming

A. Aim: stationary target
B. Aim: moving target

X. Physical Actions

A. Walk long distances
B. Run short distances
C. Push, pull, lift heavy weights
D. Throw objects
E. Sort, fold, feed by hand
F. Make coordinated movements
G. Work long hours
H. Work under adverse conditions

Figure 5.3 Hybrid Taxonomy

1. Inspect and maintain mechanical equipment/systems
2. Inspect and maintain electrical equipment/systems
3. Troubleshoot and repair electrical equipment/systems
4. Troubleshoot and repair mechanical equipment/systems
5. Operate electronic equipment
6. Operate keyboard
7. Make drawings or sketches
8. Make spatial judgments
9. Judge movement of objects
10. Pack and load
11. Construct and assemble
12. Use repetitive hand movements
13. Operate hand-held equipment
14. Operate heavy equipment
15. Drive light wheeled vehicles
16. Fire weapons
17. Make coordinated movements
18. Demonstrate physical endurance
19. Work under adverse conditions
20. Control conflicts
21. Use individual weapons
22. Execute field techniques
23. Communicate orally
24. Communicate in writing
25. Lead peers and subordinates
26. Coach and counsel peers or subordinates
27. Direct/participate in teams
28. Solve logistical, tactical, or administrative problems
29. Analyze numerical data
30. Analyze/use figural information
31. Administration/records keeping
32. Food preparation
33. Preparation for NBC engagement
34. Providing medical treatment
35. Send and receive messages
36. Operate sensor devices
37. Use explosives
38. Give first aid

Similar rating procedures were used for the Task, Activity, and Hybrid Questionnaires. Participants were asked to consider the range of duty assignments for soldiers with eighteen months experience in their particular MOS and to complete the questionnaires from this frame of reference. SMEs first rated how frequently, on a scale of 0 (never) to 5 (most often), each task and/or activity is performed by such soldiers. After providing frequency ratings for all items, participants then rated the importance of those tasks and/or activities identified as performed by soldiers with 18 months experience in the MOS (i.e., tasks with non-zero frequency ratings). Using a scale of 0 (no importance) to 5 (extremely high importance), ratings were collected for three areas of job performance: Core Technical, General Soldiering, and overall. For the Hybrid Questionnaire, SMEs provided importance and difficulty ratings simultaneously. Difficulty was rated on a scale of 1 (easiest to learn and perform) to 5 (most difficult to learn and perform). The complete scales are presented in Volume III of this report.

After completing each questionnaire, SMEs estimated the percent of the MOS performance domain which was covered by the questionnaire. Participants who indicated less than 100% of the MOS was covered were asked to suggest any items that should be added.

Validity Ratings and Rankings

The Attribute Validity Questionnaires consists of a set of 30 cognitive, psychomotor, physical, temperament, and vocational interest attributes. The attribute taxonomy is shown in Figure 5.4. Participants received a booklet, plus oral instructions, which explained the terms "attribute," "job performance," and "validity." The booklet also contained a definition and description of individuals high, average, and low on each of the 30 attributes. Job performance was divided into five areas, based on Army Project A findings (Wing, Peterson, & Hoffman, 1984): Core Technical Proficiency, General Soldiering Proficiency, Effort and Leadership, Personal Discipline, and Physical Fitness/Military Bearing. A detailed description of the development of these questionnaires appears in chapter 4 of the Phase I report (Owens-Kurtz & Peterson, 1989).

On the Attribute Validity Judgment Questionnaire, participants rated the validity of 30 attributes for performance in two job areas: Core Technical Proficiency and General Soldiering Proficiency. Twenty-two of the thirty attributes were also rated for the remaining three job areas: Effort and Leadership, Personal Discipline, and Physical Fitness/Military Bearing. (Eight vocational interest attributes were not rated against the latter three areas.) Ratings were made on a 9-point scale, from "No validity" (0) to "Extremely high validity" (8). Participants also completed an Attribute Ranking Questionnaire, on which they rank ordered the 30 attributes according to validity for overall performance in the MOS. Finally, participants were asked how thoroughly the attributes required for the rated MOS were covered by the 30 listed attributes, and were asked to make suggestions for additional attributes.

Editing and Handling of Missing Data

Task, Activity, and Hybrid Questionnaires

Two participants' responses were dropped from all analyses of the Task, Activity, and Hybrid data based on workshop leader comments and the sparsity of the data they provided. One participant seemed deficient in the English language. He had great difficulty understanding verbal and written instructions and completing the questionnaires. The other participant had no practical experience with the MOS she rated. She knew which tasks were performed, but she did not know the frequency with which they were performed or their importance.

Validity Ratings and Rankings

Descriptive statistics were computed on all available data for the Validity Ratings, with missing data treated as blanks. For the reliability analyses, data were screened to determine the number of missing values for each case. Only six participants had more than 10% of the values missing. All cases were retained and missing values were replaced with zero. This approach is conservative, since using zeros for missing data tends to underestimate the reliabilities.

Validity Ranking data were screened to determine the number of missing data for each case. Three cases had greater than 10% missing values. Upon closer inspection, it was discovered that these three cases had no data at all; they were dropped from the file. All other missing values were treated as blanks in the computation of descriptive statistics and were replaced with zeros for the reliability analyses.

Coverage of MOS Performance Domain

Task, Activity, and Hybrid Questionnaires

Table 5.1 presents a summary of the raters' evaluation of the comprehensiveness of the questionnaires in covering job content. Although no statistical analyses were conducted, the Task, Activity, and Hybrid Questionnaires seem to be essentially equal in inclusiveness, although the 94B raters indicated the Task Questionnaire was less comprehensive than either of the other two questionnaires. Raters who indicated that the questionnaire covered less than 100% of the MOS they were rating were requested to list any tasks and/or activities that had been omitted from the questionnaire. All raters were asked to make any additional comments regarding the questionnaires. Table 5.2 presents the frequency of rater comments including omitted tasks and/or activities. In comments solicited from raters, many were distracted by the fact that a large proportion of the tasks and/or activities were not relevant to their MOS. A list of omitted tasks and/or activities are in Volume II, Appendix C. Across MOS, essentially no more comments giving suggestions for changing the coverage of the questionnaire were given for the Task Questionnaire, the Activity Questionnaire, or the Hybrid Questionnaire.

Validity Ratings and Rankings

Responses to the question, "What percentage of the attributes required for performance in the MOS you are rating was covered?" are summarized in Table 5.3. The average percent covered across MOS is 91.60% (SD = 4.60), which seems to indicate more than adequate coverage. Participants were asked to suggest additional attributes if they felt the job was not adequately covered. Table 5.4 presents the number of participants who made suggestions in each MOS. Volume II, Appendix D summarizes the suggestions by attribute type (Table D-1) and by MOS (Table D-2). Twenty-six temperament attributes were suggested for addition. Several of these appear to us to be covered by the present list of attributes (e.g., self-respect/self-enthusiasm is probably part of the present attribute dominance/confidence). However, some may warrant further consideration, such as loyalty, which was mentioned by four participants.

TABLE 5.1

SUMMARY OF JUDGES' EVALUATION OF INSTRUMENTS FOR COMPREHENSIVENESS: "What percent of the MOS you are rating is covered by these tasks, activities, or tasks and activities?"

MOS	Task		Activity		Hybrid	
	Mean	N	Mean	N	Mean	N
16S	84%	78	86%	74	81%	68
19K	89%	50	93%	49	90%	50
67N	84%	50	92%	51	91%	48
76Y	82%	47	83%	45	83%	43
88M	83%	41	80%	44	77%	45
91A	79%	54	85%	54	82%	53
94B	67%	37	87%	36	83%	39
Average	84%		87%		84%	

TABLE 5.2

FREQUENCY OF COMMENTS BY MOS ON THE TASK, ACTIVITY, AND HYBRID QUESTIONNAIRES

MOS	Task	Activity	Hybrid
16S	14	16	9
19K	9	20	11
67N	17	14	15
76Y	11	8	9
88M	6	8	10
91A	22	19	21
94B	20	13	16
Total	99	98	91

TABLE 5.3

MEAN RESPONSE BY MOS TO QUESTION "WHAT PERCENTAGE OF THE ATTRIBUTES REQUIRED FOR PERFORMANCE IN THE MOS YOU ARE RATING WAS COVERED?"

MOS	Mean %
16S	87.16
19K	97.55
67N	96.54
76Y	93.57
88M	85.35
91A	91.57
94B	89.47

Average of MOS Means = 91.60

SD of MOS Means = 4.60

TABLE 5.4

NUMBER OF JUDGES WHO SUGGESTED ADDITIONAL ATTRIBUTES BY MOS

MOS	Freq
16S	14
19K	4
67N	7
76Y	5
88M	2
91A	6
94B	11
Total	49

Descriptive Statistics

Task, Activity, and Hybrid Questionnaires

For each of the seven Phase II MOS, means, standard deviations, and N's for Frequency (FRE), Core Technical Importance (CTI), General Soldiering Importance (GSI), Overall Job Importance (OJI), and Difficulty (DIF) were calculated. Note that difficulty ratings were obtained only for the Hybrid Questionnaire. Table 5.5-5.7 present the means of the CTI ratings for each of the Phase II MOS for the Task, Activity, and Hybrid Questionnaires, respectively. The complete results for means of different types ratings for the three questionnaires are included in Tables E-1 through E-21 of Volume II, Appendix E.

Recall that when the questionnaires were administered, participants provided frequency ratings for all items and importance ratings for only those items with non-zero frequency ratings. In calculating descriptive statistics, importance ratings associated with zero frequency ratings were set to zero rather than treated as missing. Given the rating procedure employed, a zero frequency rating implies a zero importance rating and these zero ratings should be considered when examining the mean importance of items. Coding non-rated items as missing, rather than zero, would result is an overestimation of the importance of those items.

Validity Ratings and Rankings

For each MOS, attribute means and standard deviations were computed for validity judgments against the five job performance areas and for validity rankings against overall performance. Table 5.8 presents the mean validity ratings for Core Technical Proficiency for the seven Phase II MOS. Volume II, Appendix F presents the full results, and includes Phase I MOS for comparison when available. (One attribute, closure, was included in Phase I but not in Phase II. In order to make the data comparable, closure was excluded from Phase I data for all analyses in this report.)

TABLE 5.5

TASK QUESTIONNAIRE MEAN CORE TECHNICAL IMPORTANCE RATINGS FOR PHASE II MOS

Task Categories				MOS			
	16S N*=89	19K N=53	67N N=58	76Y N=49	88M N=52	91A N=58	94B N=42
Perform operator maint chks and services	4.14	4.46	4.16	3.18	4.56	2.71	3.05
Perform op checks and services on weapons	4.01	4.42	2.41	3.42	3.19	2.28	2.52
Troubleshoot mechanical systems	1.39	3.35	4.38	0.88	2.10	1.19	1.00
Repair weapons	1.06	2.58	0.81	2.74	0.38	0.76	0.51
Repair mechanical systems	1.62	3.42	4.29	0.96	2.06	1.29	1.40
Troubleshoot weapons	1.56	3.37	1.50	2.42	1.18	0.91	1.00
Install electronic components	2.66	3.04	2.79	0.68	0.73	1.28	0.37
Inspect electrical systems	0.69	2.23	3.19	0.32	0.86	0.89	0.51
Inspect electronic systems	0.90	2.17	1.86	0.32	0.27	0.62	0.23
Repair electrical systems	0.50	1.63	2.36	0.16	0.51	0.47	0.26
Repair electronic components	0.32	1.17	1.66	0.30	0.27	0.38	0.21
Pack and load materials	2.45	2.48	2.84	3.30	3.65	2.02	2.67
Prepare parachutes	0.17	0.10	0.14	0.30	0.02	0.26	0.07
Prepare equip and supplies for air drop	0.23	0.14	0.36	1.20	0.65	0.38	0.30
Operate power excavating equipment	0.08	0.15	0.03	0.24	0.06	0.19	0.02
Operate wheeled vehicles	4.23	2.31	2.37	3.36	4.58	2.78	2.47
Operate track vehicles	1.28	4.38	0.09	0.36	0.33	1.76	0.05
Operate boats	0.10	0	0	0.08	0.14	0.17	0.02
Operate lifting, loading, & grading equip	0.07	0.08	0.53	0.72	0.39	0.24	0.07
Paint	1.27	1.73	1.83	1.22	1.77	0.67	1.44
Install wire and cables	2.18	1.88	1.10	0.60	0.49	0.50	0.19
Repair plastic and fiberglass	0.23	0.04	1.70	0.18	0.21	0.17	0
Repair metal	0.27	1.40	1.97	0.28	0.41	0.29	0.19
Assemble steel structures	0.36	0.50	0.14	0.24	0.14	0.19	0
Install pipe assemblies	0.04	0.17	0.52	0.16	0.18	0.14	0.24
Construct wooden bldgs and other structures	0.22	0.08	0.24	0.24	0.06	0.19	0.07
Construct masonry bldgs and structures	0.17	0.10	0.03	0.10	0.02	0.07	0
Operate gas and electric powered equip	1.33	1.15	2.50	1.30	0.94	1.14	2.37
Select, layout, & clean med/den equipment	0.04	0.06	0.10	0.14	0.02	3.34	0.12
Use audiovisual equipment	0.79	0.73	0.53	0.92	0.27	1.02	0.23
Reproduce printed material	0.49	0.79	0.52	1.90	0.39	0.78	0.63
Operate electronic equipment	2.12	2.29	1.59	1.04	0.53	1.14	0.26
Operate radar	0.34	0.12	0.12	0.18	0.04	0.14	0
Operate computer hardware	0.32	0.73	0.55	2.67	0.20	0.83	0.21
Cook	0.10	0.13	0.05	0.14	0.14	0.52	4.80
Perform medical laboratory procedures	0	0	0	0.10	0.02	1.97	0
Conduct land surveys	1.06	1.10	0.71	0.50	0.60	1.26	0.40
Provide medical or dental treatment	0.03	0.10	0.28	0.28	0.04	4.49	0.09
Sketch maps, overlaps, or range cards	3.13	3.42	1.10	1.20	1.85	1.35	0.53
Produce technical drawings	0.06	0.31	0.26	0.22	0.04	0.14	0.02
Draw maps and overlays	1.46	1.31	0.26	0.51	0.49	0.29	0.05
Draw illustrations	0.37	0.38	0.57	0.43	0.24	0.36	0.14
Type	0.19	0.88	0.95	3.74	0.46	1.46	1.98
Prepare technical forms and documents	1.00	1.94	4.00	4.04	1.78	2.21	2.14

TABLE 5.5 (CONTINUED)

TASK QUESTIONNAIRE MEAN CORE TECHNICAL IMPORTANCE RATINGS ACROSS MOS

				MOS			
Task Categories	16S	19K	67N	76Y	88M	91A	94B
	N*=89	N=53	N=58	N=49	N=52	N=58	N=42
Record, file, and dispatch information	0.86	1.08	2.72	4.29	1.51	2.52	2.02
Receive, store, & issue supp, equip, etc	0.75	1.04	2.29	4.57	1.65	2.19	3.43
Use hand and arm signals	2.67	3.77	3.05	1.38	3.17	1.59	1.05
Read tech manuals, field manuals, regs etc	3.63	3.88	4.81	3.60	3.94	2.93	3.81
Use maps	4.17	3.92	2.79	2.44	4.02	2.93	2.33
Send and receive radio messages	4.17	3.87	2.33	2.00	2.29	2.59	0.72
Give short oral reports	3.57	3.65	2.00	2.26	2.40	2.50	1.05
Receive clients, patients, guests	0.04	0.35	0.26	0.44	0.10	3.43	0.40
Give directions and instructions	2.90	2.88	2.90	2.72	2.83	3.53	2.88
Write and deliver presentations	0.48	0.58	0.53	0.78	0.24	1.07	0.42
Interview	0.29	0.46	0.40	0.24	0.39	2.86	0.35
Provide counseling	1.12	1.12	1.19	0.82	0.80	1.86	1.12
Write documents and correspondence	0.46	0.65	0.79	1.80	0.51	1.40	0.58
Decode data	3.39	2.77	1.00	0.84	1.12	1.21	0.23
Analyze electronic signals	0.75	0.94	0.34	0.06	0.12	0.19	0.19
Analyze weather conditions	0.93	1.38	0.60	0.20	0.61	0.71	0.37
Order equipment and supplies	1.27	1.67	3.14	4.35	0.90	2.59	1.93
Estimate time and cost of maint ops	0.14	0.44	1.71	1.14	0.35	0.59	0.47
Plan placement/use of tactical equip	2.74	2.60	1.28	1.08	1.29	1.03	0.72
Translate foreign languages	0.01	0.17	0.03	0.18	0.33	0.36	0.02
Analyze intelligence data	1.25	1.16	0.26	0.16	0.33	0.57	0.12
Control money	0.19	0.40	0.10	1.70	0.33	0.47	2.12
Determine firing data-indirect weapons	0.69	1.19	0.22	0.16	0.35	0.24	0.14
Compute statistics/other math	0.04	0.73	1.10	1.00	0.35	1.62	1.56
Provide programming and DP support	0.09	0.04	0.12	0.72	0.08	0.36	0.19
Control air traffic	0.11	0.04	0.28	0.14	0.02	0.47	0
Use hand grenades	2.30	2.47	0.98	1.31	1.63	0.67	0.81
Protect against NBC hazards	3.82	4.35	2.86	2.88	3.60	3.19	2.86
Handle demolitions or mines	1.40	2.60	0.44	0.60	0.41	0.24	0.09
Engage in hand-to-hand combat	1.27	1.56	0.53	0.63	0.69	0.62	0.53
Fire individual weapons	3.76	3.88	2.32	2.70	2.98	2.22	2.24
Control individuals and crowds	1.51	2.02	1.09	0.81	1.37	1.31	0.63
Customs and laws of war	2.25	2.46	1.25	1.63	1.90	2.55	1.05
Navigate	3.99	3.71	1.93	1.75	3.06	2.57	1.44
Survive in the field	4.06	3.85	2.14	1.94	2.85	2.47	2.00
Move and react in the field	3.60	3.10	1.75	1.69	1.75	2.29	1.42
Load and unload field artil/tank guns	0.17	4.31	0.04	0.12	0.27	0.19	0.02
Fire heavy direct fire weapons	0.50	4.46	0.02	0.06	0.10	0.12	0.02
Prepare heavy weapons for tactical use	0.53	2.10	0	0.10	0.10	0.10	0
Place & camoufl tactical equip and mat	3.03	3.75	1.73	1.65	1.54	0.90	1.07
Fire indirect fire weapons	0.20	0.54	0	0.20	0.17	0.10	0.02
Give first aid	3.45	3.50	2.30	2.44	2.98	4.57	2.91

TABLE 5.5 (CONTINUED)

TASK QUESTIONNAIRE MEAN CORE TECHNICAL IMPORTANCE RATINGS ACROSS MOS

Task Categories			MOS				
	16S N*=89	19K N=53	67N N=58	76Y N=49	88M N=52	91A N=58	94B N=42
Detect and identify targets	4.49	4.29	2.04	1.06	1.51	0.97	0.84
Plan operations	2.22	1.33	1.44	1.14	0.73	1.12	0.98
Direct/lead teams	2.64	1.96	1.39	1.02	0.45	0.98	0.42
Monitor/inspect	2.49	1.98	2.79	2.16	1.39	2.21	2.23
Lead	3.02	2.44	3.07	2.04	1.94	2.34	2.19
Act as a model	3.39	2.73	3.61	2.76	2.56	3.02	3.30
Counsel	2.69	2.10	2.53	1.92	1.61	2.24	2.30
Communicate	3.44	2.65	3.28	3.00	2.40	2.84	2.56
Train	2.69	1.92	2.60	2.18	1.84	2.55	2.35
Personnel Administration	1.48	1.21	1.54	2.28	0.86	1.71	1.70

*NOTE: N = total number of participants from each MOS. Due to missing data, some table entries are based on smaller samples.

TABLE 5.6

JOB ACTIVITY MEAN CORE TECHNICAL IMPORTANCE RATINGS FOR PHASE II MOS

Job Activities				MOS			
	16S N*=89	19K N=53	67N N=58	76Y N=49	88M N=52	91A N=58	94B N=42
Work in a team	4.15	3.85	3.86	2.94	2.26	3.52	3.79
Lead a team	3.75	2.62	3.05	1.94	1.79	2.34	2.38
Support/advise peers	3.19	2.62	3.18	2.50	2.07	3.09	2.88
Support/advise subordinates	3.13	2.50	3.36	2.44	2.21	2.50	2.69
Coach peers	3.38	2.92	2.93	2.30	2.05	2.88	2.69
Coach subordinates	3.49	2.73	3.32	2.58	2.07	2.79	2.67
Make oral reports (to individuals)	2.86	2.40	3.02	2.34	2.19	3.20	1.86
Make oral reports (to groups)	1.88	1.38	1.89	1.52	1.23	2.04	1.52
Relay oral instructions	3.25	3.33	3.52	2.76	2.47	3.14	2.90
Interview	1.28	0.85	1.84	1.16	0.55	2.71	1.33
Record information	2.81	2.54	3.98	3.27	2.86	3.55	2.69
Write brief messages	1.98	2.04	2.43	2.70	1.23	2.79	1.74
Write longer reports	0.82	0.73	0.88	1.34	0.43	1.09	0.64
Monitor/interpret verbal messages	2.91	2.98	2.66	2.26	1.19	2.66	1.64
Recall verbal information	2.97	3.29	3.20	2.64	2.33	3.39	2.48
Monitor/interpret numerical information	2.01	2.29	2.32	2.16	1.00	2.31	1.60
Recall numerical information	2.00	2.40	2.54	2.14	1.33	2.65	1.98
Monitor/interpret figural information	2.56	2.37	2.68	1.36	1.27	2.04	1.05
Recall figural information	2.76	2.19	2.18	1.08	1.55	2.00	0.88
Follow oral directions	3.90	4.10	4.13	3.82	3.84	3.89	4.07
Follow written directions	3.66	3.81	4.61	3.82	4.14	3.98	4.24
Judge size and distance	3.94	3.54	2.54	1.38	3.40	1.79	1.56
Judge location	4.13	3.73	2.36	1.92	3.70	2.54	1.84
Judge paths of moving objects	3.70	2.96	2.68	1.18	3.84	1.57	1.30
Solve electrical system problems	1.33	2.37	3.50	0.46	1.23	0.96	0.84
Solve mechanical system problems	2.11	3.13	4.34	1.10	2.47	1.70	1.74
Solve logistical problems	1.22	1.31	2.25	3.28	0.95	1.52	1.21
Solve tactical maneuver problems	2.68	2.15	1.29	0.64	0.82	1.02	0.74
Solve administrative problems	1.20	1.35	2.24	2.90	0.95	1.55	2.31
Solve leadership problems	2.11	1.76	2.07	1.78	1.07	1.86	1.83
Solve medical problems	0.74	0.94	0.89	0.44	0.27	3.48	1.02
Solve communication problems	2.09	1.65	1.73	1.55	0.95	1.71	1.59
Operate precision hand-held equipment	1.36	1.27	3.84	0.60	0.59	3.36	1.24
Operate hand-held tools	2.60	3.65	4.57	1.76	3.31	2.38	2.60
Operate hand-held power equipment	0.92	2.04	3.59	0.92	1.43	0.95	1.57
Operate larger power equipment	0.14	1.54	1.04	0.92	0.73	0.30	0.37
Operate full keyboard	0.22	0.56	1.00	3.38	0.43	1.20	1.52
Operate numeric keyboard	0.16	1.52	0.66	1.98	0.20	0.50	0.90
Adjust device using one limb	2.46	3.40	2.96	1.20	3.60	2.16	2.05
Adj control device using multiple limbs	2.56	3.62	3.16	1.10	3.57	2.32	1.60
Drive tracked vehicle	1.18	4.54	0.07	0.46	0.63	1.75	0.09
Drive heavy wheeled vehicle	0.99	2.10	1.04	1.59	4.29	1.57	1.34
Drive light wheeled vehicle	4.25	2.38	2.46	3.18	4.36	2.73	1.95
Aim:stationary target	3.42	4.27	2.21	1.76	2.49	1.93	1.85

TABLE 5.6 (CONTINUED)

JOB ACTIVITY MEAN CORE TECHNICAL IMPORTANCE RATINGS FOR PHASE II MOS

				MOS			
Job Activities	16S	19K	67N	76Y	88M	91A	94B
	N*=89	N=53	N=58	N=49	N=52	N=58	N=42
Aim:moving target	4.47	4.21	2.09	0.98	1.77	1.23	1.02
Walk long distances	3.06	2.12	1.78	1.59	1.56	2.59	1.67
Run short distances	3.57	2.81	2.27	2.18	2.60	2.88	2.29
Push, pull, lift heavy weights	2.83	3.54	2.82	2.50	2.38	3.21	2.81
Throw objects	1.63	1.96	0.86	0.96	1.14	0.86	0.63
Sort, fold, feed by hand	0.65	1.73	0.89	2.06	0.69	1.43	2.43
Make coordinated movements	2.84	3.00	3.02	1.66	2.81	3.09	2.83
Work long hours	3.33	3.88	3.25	3.00	3.24	3.32	4.00
Work under adverse conditions	3.56	4.02	3.29	2.24	3.36	3.39	3.60

*NOTE: N = total number of participants from each MOS. Due to missing data, some table entries are based on smaller samples.

TABLE 5.7

HYBRID MEAN CORE TECHNICAL IMPORTANCE RATINGS FOR PHASE II MOS

Hybrid Elements				MOS			
	16S N*=89	19K N=53	67N N=58	76Y N=49	88M N=52	91A N=58	94B N=42
Insp & maint mechanical equip/systems	4.10	4.25	3.91	2.32	4.45	2.77	3.43
Insp & maint electrical equip/systems	3.44	3.79	3.58	0.92	1.94	1.55	1.28
Tblsht & repair electrical equip/sys	1.01	2.27	3.05	0.40	0.68	0.60	0.33
Tblsht & repair mechanical equip/sys	1.68	2.98	4.42	0.72	1.68	1.09	1.42
Operate electronic equipment	1.89	2.87	2.14	2.26	0.43	1.07	0.56
Operate keyboard	0.16	1.13	1.00	3.56	0.45	1.04	1.49
Make drawings or sketches	2.99	2.90	1.19	1.00	2.26	1.22	0.60
Make spatial judgments	3.38	3.10	2.00	1.58	2.93	1.81	1.33
Judge movement of objects	3.40	3.12	2.32	0.81	2.62	1.14	0.91
Pack and load	2.58	2.62	2.49	2.82	3.36	2.57	2.53
Construct and assemble	0.91	0.54	0.89	0.45	0.39	0.88	0.53
Use repetitive hand movements	1.62	1.60	3.11	1.69	1.32	2.33	2.53
Operate hand-held equipment	2.74	3.35	4.25	1.66	3.20	2.02	2.65
Operate heavy equipment	1.18	4.23	1.25	1.22	4.20	1.43	0.77
Drive light wheeled vehicles	3.93	2.08	2.21	3.24	4.43	3.05	2.51
Fire weapons	3.30	4.48	2.46	2.31	2.72	1.26	1.77
Make coordinated movements	3.07	3.04	2.56	1.84	2.30	2.28	1.49
Demonstrate physical endurance	3.52	3.48	2.44	2.68	3.37	3.31	3.30
Work under adverse conditions	3.39	3.87	2.79	2.28	3.13	2.98	3.23
Control conflicts	1.06	1.62	0.79	0.75	0.62	1.46	0.79
Use individual weapons	3.80	3.31	2.12	2.85	3.24	2.30	2.23
Execute field techniques	3.72	3.87	2.02	2.13	3.30	2.40	2.98
Communicate orally	3.07	2.87	3.07	2.77	2.48	3.33	2.35
Communicate in writing	1.85	1.79	2.19	2.77	1.96	2.96	1.58
Lead peers or subordinates	2.94	2.73	3.16	2.56	2.45	2.26	2.21
Coach & counsel peers/subordinates	2.67	2.35	2.71	2.44	1.72	2.09	2.05
Direct/participate in teams	3.51	3.58	3.41	2.44	2.51	2.81	3.05
Solve logistic/tactic/admin problems	1.39	1.08	1.45	2.82	0.74	1.52	1.19
Analyze numerical data	0.66	1.00	1.50	1.40	0.28	1.14	0.63
Analyze/use figural information	3.19	2.17	1.71	1.56	1.43	1.38	0.86
Administration/records keeping	1.21	1.31	3.25	4.26	2.13	2.74	2.47
Food preparation	0.35	0.15	0.28	0.26	0.19	0.56	4.44
Preparation for NBC engagement	3.34	3.83	2.58	2.10	2.91	2.68	2.52
Providing medical treatment	0.47	0.58	0.35	0.52	0.34	4.19	0.19
Send and receive messages	3.55	3.33	2.09	2.24	1.81	2.47	1.12
Operate sensor devices	1.08	1.00	0.21	0.22	0.11	0.43	0.23
Use explosives	1.86	2.94	1.05	1.06	1.81	0.39	0.40
Give first aid	3.11	3.73	2.64	2.22	2.85	4.64	2.84

*NOTE: N = total number of participants from each MOS. Due to missing data, some table entries are based on smaller samples.

TABLE 5.8

ATTRIBUTE MEAN VALIDITY RATINGS FOR

CORE TECHNICAL PROFICIENCY FOR PHASE II MOS

		MOS					
	16S	19K	67N	76Y	88M	91A	94B
ATTRIBUTE	N*=89	N=53	N=58	N=49	N=52	N=58	N=42
Verbal Ability	5.000	5.019	5.724	5.388	4.731	6.000	4.381
Reasoning	4.843	4.830	5.793	4.837	4.519	5.707	4.595
Number Ability	3.674	4.075	4.828	5.347	3.981	5.224	4.976
Spatial Ability	4.978	4.528	5.690	3.833	4.135	4.138	3.095
Mental Information Processing	5.775	5.736	5.517	5.061	5.173	6.034	4.571
Perceptual Speed & Accuracy	5.584	5.283	5.569	4.796	4.769	5.552	4.119
Memory	6.101	6.000	5.845	5.204	5.462	5.983	4.952
Mechanical Comprehension	4.573	6.113	7.121	3.020	5.923	4.052	3.690
Eye-Limb Coordination	5.652	5.849	5.914	4.042	6.038	4.966	4.857
Precision	6.528	5.906	5.517	3.313	4.846	4.672	4.190
Movement Judgment	6.472	5.377	4.586	2.837	5.750	3.638	2.738
Hand & Finger Dexterity	5.202	5.340	6.586	3.771	4.922	5.345	5.071
Physical Strength	5.213	5.113	4.672	4.673	5.808	4.741	5.286
Physical Endurance	5.472	4.830	4.431	4.388	5.385	5.259	5.071
Balance and Flexibility	4.978	4.377	5.224	3.396	4.423	4.190	4.048
Involvement in Athletics	3.573	3.472	2.983	2.857	3.731	3.121	2.976
Work Orientation	5.348	5.925	6.793	5.490	5.442	5.828	5.881
Sociability	3.573	3.906	4.138	3.939	3.692	4.690	4.310
Cooperation/Stability	4.539	4.830	5.276	5.143	4.577	5.569	5.262
Energy	4.955	5.075	5.741	5.122	5.096	5.431	5.500
Conscientiousness	5.056	5.019	6.052	5.224	5.596	5.724	5.167
Dominance/Confidence	4.727	5.019	4.966	4.939	5.058	4.914	4.714
Interest in Using Tools & Machines	4.506	5.750	7.086	3.184	6.173	3.810	3.619
Interest in Rugged Activities	4.865	5.327	4.086	3.061	4.481	3.552	3.024
Interest in Protective Services	4.112	3.654	3.862	4.061	3.538	4.500	3.452
Interest in Technical Activities	3.899	4.231	5.965	3.592	3.673	4.414	3.571
Interest in Science	2.775	2.769	4.276	2.041	2.308	5.552	2.476
Interest in Leadership	5.112	4.942	4.879	4.531	4.731	4.810	4.595
Interest in Artistic Activities	1.910	1.808	2.121	1.755	2.135	2.983	3.786
Interest in Efficiency & Organization	4.281	4.385	5.397	5.918	4.481	5.000	5.476

*NOTE: N = total number of participants from each MOS. Due to missing data, some table entries are based on smaller samples.

Reliability Analyses

Task, Activity, and Hybrid Questionnaires

Approach

Variance component analyses and generalizability coefficients (e.g., Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson, Webb, & Rowley, 1989) were calculated to examine simultaneously: (a) mean differences in ratings of various rater groups, (b) reliability across rater groups, and (c) reliability within rater groups.

For each rating scale (FRE, CTI, GSI, OJI, and DIF) within the three questionnaires (Task, Activity, and Hybrid) and for each of the seven Phase II MOS (16S, 19K, 67N, 76Y, 88M, 91A, and 94B), four generalizability models were run; for convenience, these are labeled Type A, B, C, and D analyses:

(1) Variance components for Type A analysis included item, rank, command, and rater nested within rank and command. Type A analyses were run for each of the seven MOS.

(2) Type B analysis included item, rank, and rater nested within rank as facets. Separate analyses were run for TRADOC and FORSCOM resulting in 14 MOS-by- command combinations.

(3) Type C analysis included item, command, and rater nested within command as facets. Separate analyses were run for NCO, Officer, and Civilian (for 2 MOS only) resulting in 16 MOS-by-rank combinations.

(4) Facets for Type D analysis included item and rater. Type D analyses were run for 28 MOS-by-rank-by-command combinations.

Type A, B, and C analyses each required two computer runs. The first run included only item, rater, and the interaction. The second run included item, rank, and/or command and the interactions, but not rater. Rater, nested within rank and/or command, was calculated by subtracting variance attributed to rank and/or command obtained in Run 2 from the rater and variance estimate obtained in Run 1. Similarly, item- by-rater, nested within rank and/or command, was calculated by subtracting variance attributed to rank and/or command obtained in Run 2 from the item-by-rater variance estimate obtained in Run 1. Variance attributed to item is the reliable source of variance, and when divided by total variance, provides estimates of single-rater reliability across all rater types.

For Type A, B, and C analyses, estimates were also made of single-rater reliability within rater types. These estimates were calculated after removing variance due to rater type from total variance. For example, for Type C analyses of FORSCOM and TRADOC raters, single-rater reliabilities within rank were computed by removing rank and its interactions from the total variance estimate. Item variance was then divided by adjusted total variance to estimate single-rater reliability within rank. No variance component estimates were calculated for the entire sample due to computer time constraints.

Results

Tables 5.9, 5.10, and 5.11 present summary single-rater reliability estimates for the Task, Activity, and Hybrid Questionnaires, respectively. Detailed variance component estimates from which these reliabilities were computed are presented in Volume II, Appendix G. Single-rater reliability estimates for Type A, B, and C analyses can be found in Volume II, Appendix H.

Differences in reliability among rater groups. If rank, command, or rank-by-command effects were large, the change in single-rater reliabilities from overall to within rank, overall to within command, or overall to within rank and command would be pronounced. This occurs only twice: (1) for the 67N on the Task Questionnaire, and (2) for the 76Y on the Hybrid Questionnaire. These effects for 67N and 76Y hold regardless of the rating scale. Examination of the Type A detailed variance components tables in Appendix G (Tables G-1 through G-5) suggests that the 67N effect is

TABLE 5.9

TASK QUESTIONNAIRE: SUMMARY OF SINGLE-RATER RELIABILITY ESTIMATES

		TRADOC			FORSCOM			COMBINED COMMAND			
		NCO	Officer	Total	NCO	Officer	Total	NCO	Officer	Civilian	Total
16S	N	11	9	22	42	26	68	53	35	2*	90
	FRE	.57	.63	.59	.51	.61	.55	.52	.63		.57
	CTI	.56	.60	.58	.49	.57	.53	.52	.60		.46
	GSI	.58	.61	.60	.49	.59	.54	.54	.62		.57
	OJI	.58	.49	.59	.49	.58	.54	.54	.61		.56
19K	N	10	4	22	16	15	31	26	19	8*	53
	FRE	.48	.75	.56	.49	.56	.53	.49	.59	.53	.54
	CTI	.51	.66	.55	.53	.57	.55	.52	.59	.53	.55
	GSI	.51	.60	.52	.50	.53	.52	.51	.55	.42	.53
	OJI	.51	.69	.55	.51	.57	.54	.51	.59	.47	.55
67N	N	23	8	32	11	15	26	34	23	1*	58
	FRE	.50	.56	.52	.56	.62	.61	.52	.58		.58
	CTI	.45	.52	.47	.47	.61	.57	.47	.57		.54
	GSI	.48	.47	.47	.46	.61	.56	.49	.53		.55
	OJI	.51	.46	.49	.48	.63	.58	.52	.57		.56
76Y	N	12	8	21	16	13	29	28	21	1*	50
	FRE	.46	.52	.51	.53	.52	.52	.51	.53		.53
	CTI	.41	.52	.46	.47	.49	.47	.45	.51		.73
	GSI	.38	.46	.41	.46	.44	.44	.43	.45		.44
	OJI	.40	.52	.46	.47	.47	.47	.45	.49		.47

TABLE 5.9 continued

		TRADOC			FORSCOM			COMBINED COMMAND			
		NCO	Officer	Total	NCO	Officer	Total	NCO	Officer	Civilian	Total
88M	N	13	5	23	17	9	26	30	14	4*	49
	FRE	.47	.50	.50	.50	.49	.51	.50	.51	.53	.51
	CTI	.41	.53	.45	.50	.53	.53	.52	.53	.47	.50
	GSI	.39	.41	.40	.50	.49	.50	.45	.46	.44	.46
	OJI	.41	.46	.42	.47	.53	.51	.44	.51	.49	.48
91A	N	15	10	25	15	18	33	30	28	0	58
	FRE	.44	.42	.43	.41	.57	.53	.42	.50		.50
	CTI	.38	.41	.39	.38	.51	.45	.37	.47		.43
	GSI	.43	.30	.38	.40	.57	.51	.41	.48		.47
	OJI	.43	.35	.41	.40	.57	.51	.41	.38		.48
94B	N	9	6	15	17	12	29	26	18	0	44
	FRE	.38	.41	.41	.41	.50	.44	.41	.49		.44
	CTI	.39	.38	.40	.42	.48	.44	.41	.45		.43
	GSI	.37	.39	.39	.40	.49	.43	.40	.46		.42
	OJI	.39	.45	.43	.42	.47	.44	.42	.47		.44
Mean	FRE	.47	.54	.50	.49	.55	.53	.48	.55	.53	.52
	CTI	.44	.52	.47	.47	.54	.51	.47	.53	.50	.52
	GSI	.45	.46	.45	.46	.53	.50	.46	.51	.43	.49
	OJI	.46	.46	.48	.46	.55	.51	.47	.52	.48	.51

* All from TRADOC.

TABLE 5.10

ACTIVITY QUESTIONNAIRE: SUMMARY OF SINGLE-RATER RELIABILITY ESTIMATES

		TRADOC			FORSCOM			COMBINED COMMAND			
		NCO	Officer	Total	NCO	Officer	Total	NCO	Officer	Civilian	Total
16S	N	10	8	20	42	26	68	52	34	2*	88
	FRE	.44	.52	.49	.42	.53	.45	.43	.55		.49
	CTI	.42	.46	.47	.39	.46	.40	.41	.48		.44
	GSI	.38	.46	.45	.37	.46	.42	.39	.49		.45
	OJI	.40	.47	.46	.40	.49	.43	.43	.51		.47
19K	N	10	4	22	16	15	31	26	19	8*	53
	FRE	.32	.48	.40	.25	.44	.32	.28	.47	.47	.38
	CTI	.32	.38	.37	.23	.41	.31	.29	.43	.42	.37
	GSI	.35	.30	.34	.25	.33	.29	.33	.33	.31	.34
	OJI	.32	.37	.37	.24	.36	.28	.27	.39	.40	.35
67N	N	23	8	31	11	14	25	34	22	0	56
	FRE	.31	.52	.35	.28	.41	.37	.29	.48		.36
	CTI	.30	.50	.34	.30	.45	.40	.30	.49		.38
	GSI	.28	.39	.32	.25	.35	.29	.28	.36		.32
	OJI	.28	.44	.31	.29	.37	.35	.28	.40		.34
76Y	N	12	8	21	16	13	29	28	21	1*	50
	FRE	.20	.36	.26	.22	.34	.24	.21	.38		.25
	CTI	.20	.37	.26	.18	.33	.22	.19	.37		.24
	GSI	.16	.32	.21	.19	.22	.20	.19	.26		.20
	OJI	.19	.31	.23	.19	.27	.20	.19	.30		.22

TABLE 5.10 continued

		TRADOC			FORSCOM			COMBINED COMMAND			
		NCO	Officer	Total	NCO	Officer	Total	NCO	Officer	Civilian	Total
88M	N	13	5	23	12	9	21	25	14	4*	44
	FRE	.40	.48	.43	.37	.56	.44	.38	.50	.48	.45
	CTI	.35	.48	.39	.35	.47	.39	.34	.47	.44	.39
	GSI	.29	.41	.32	.32	.34	.32	.29	.37	.34	.31
	OJI	.34	.44	.37	.32	.45	.36	.32	.44	.35	.36
91A	N	15	10	25	14	17	31	29	27	0	56
	FRE	.31	.33	.30	.25	.33	.28	.26	.32		.29
	CTI	.21	.39	.26	.22	.36	.27	.22	.34		.26
	GSI	.28	.20	.24	.23	.37	.29	.24	.48		.27
	OJI	.25	.28	.25	.24	.36	.29	.24	.30		.27
94B	N	9	6	15	17	12	29	26	18	0	44
	FRE	.20	.14	.19	.36	.41	.36	.28	.31		.29
	CTI	.24	.07	.15	.34	.43	.36	.28	.23		.27
	GSI	.25	.10	.17	.32	.38	.33	.28	.25		.27
	OJI	.23	.08	.17	.30	.37	.32	.26	.24		.25
Mean	FRE	.31	.40	.35	.31	.43	.35	.30	.43	.48	.36
	CTI	.29	.38	.32	.29	.42	.34	.29	.40	.43	.34
	GSI	.28	.31	.29	.28	.35	.31	.29	.36	.33	.31
	OJI	.29	.34	.31	.28	.38	.32	.28	.37	.38	.32

* All from TRADOC.

TABLE 5.11

HYBRID QUESTIONNAIRE: SUMMARY OF SINGLE-RATER RELIABILITY ESTIMATES

		TRADOC			FORSCOM			COMBINED COMMAND			
		NCO	Officer	Total	NCO	Officer	Total	NCO	Officer	Civilian	Total
16S	N	10	9	21	42	26	68	52	35	2*	89
	FRE	.45	.48	.45	.44	.50	.47	.44	.51		.47
	CTI	.40	.40	.39	.39	.43	.41	.40	.43		.41
	GSI	.41	.43	.42	.42	.45	.44	.43	.45		.45
	OJI	.44	.41	.43	.43	.46	.44	.44	.45		.46
	DIF	.04	.14	.28	.12	.20	.32	.31	.34		.33
19K	N	10	4	22	16	15	31	26	19	8*	53
	FRE	.37	.40	.40	.37	.49	.42	.35	.49	.50	.42
	CTI	.39	.44	.41	.39	.48	.43	.39	.51	.42	.43
	GSI	.44	.39	.39	.41	.43	.43	.42	.41	.33	.41
	OJI	.41	.38	.41	.41	.47	.43	.41	.45	.42	.42
	DIF	.10	.09	.24	.05	.25	.29	.24	.27	.31	.28
67N	N	22	8	31	11	15	26	33	23	1*	57
	FRE	.30	.39	.32	.37	.52	.46	.34	.48		.39
	CTI	.29	.41	.30	.41	.49	.45	.34	.47		.38
	GSI	.32	.30	.33	.34	.43	.39	.33	.38		.37
	OJI	.32	.29	.32	.37	.50	.43	.35	.44		.39
	DIF	.15	.39	.25	.18	.30	.39	.31	.37		.33

TABLE 5.11 continued

		TRADOC			FORSCOM			COMBINED COMMAND			
		NCO	Officer	Total	NCO	Officer	Total	NCO	Officer	Civilian	Total
76Y	N	12	8	21	16	13	29	28	21	1*	50
	FRE	.26	.46	.32	.29	.51	.37	.29	.52		.39
	CTI	.19	.46	.28	.25	.45	.33	.24	.48		.35
	GSI	.22	.39	.29	.26	.43	.34	.25	.43		.36
	OJI	.24	.41	.31	.27	.43	.35	.28	.45		.37
	DIF	.07	.18	.19	.12	.09	.27	.17	.34		.26
88M	N	13	5	23	19	5	24	32	10	4*	47
	FRE	.43	.53	.46	.46	.44	.47	.44	.50	.56	.47
	CTI	.39	.46	.42	.40	.50	.47	.39	.50	.54	.44
	GSI	.33	.42	.37	.41	.53	.47	.37	.49	.38	.42
	OJI	.36	.44	.39	.40	.51	.46	.38	.49	.39	.42
	DIF	.08	.27	.27	.07	.31	.34	.28	.35	.36	.31
91A	N	15	11	26	15	17	32	30	28	0	58
	FRE	.49	.28	.41	.37	.44	.41	.42	.37		.42
	CTI	.37	.23	.30	.32	.45	.37	.33	.35		.34
	GSI	.37	.19	.29	.32	.43	.37	.33	.34		.34
	OJI	.40	.21	.32	.35	.44	.39	.36	.33		.36
	DIF	.08	.21	.14	.04	.26	.24	.25	.17		.21

TABLE 5.11 continued

		TRADOC			FORSCOM			COMBINED COMMAND			
		NCO	Officer	Total	NCO	Officer	Total	NCO	Officer	Civilian	Total
94B	N	9	6	15	17	12	29	26	18	0	44
	FRE	.02	.30	.27	.32	.38	.37	.31	.36		.34
	CTI	.29	.31	.30	.34	.33	.34	.33	.34		.33
	GSI	.23	.27	.24	.32	.42	.36	.29	.38		.32
	OJI	.30	.28	.37	.34	.37	.36	.33	.34		.34
	DIF	.10	.15	.14	.06	.14	.26	.25	.16		.23
Mean	FRE	.33	.41	.38	.37	.47	.42	.37	.46	.53	.41
	CTI	.33	.39	.34	.36	.45	.40	.35	.44	.48	.38
	GSI	.33	.34	.33	.35	.45	.40	.35	.41	.36	.38
	OJI	.35	.35	.36	.37	.45	.41	.36	.42	.41	.39
	DIF	.09	.20	.22	.09	.22	.30	.26	.29	.34	.28

* All from TRADOC.

due to rank-by-command and rank-by-command-by-task interactions. The 76Y effect is due to a rank-by-command interaction. By and large, rater groups agree with each other.

Officers are consistently more reliable (i.e., agree with each other) than NCOs. However, the difference is not large and could easily be offset by the practicality of obtaining NCO raters. NCOs are easier to obtain, and their attendance at workshops was higher. They tend to be closer to the daily work of enlisted soldiers.

There is no apparent difference in the agreement among FORSCOM raters compared to the agreement among TRADOC raters.

Differences in reliability among questionnaires and scales. Single-rater reliabilities for the Task Questionnaire are higher than those for the Activity and Hybrid Questionnaires. For 76Y, 91A, and 94B, the difference between the Task and Activity Questionnaires is very noticeable. The difficulty rating, which appeared only on the Hybrid Questionnaire, shows the least reliability of any scale on any questionnaire. We speculate that because the each of the hybrid components may cover a wide range of performance, SMEs might have focussed on different aspects of job performance when making the difficulty of judgments.

Summary and Conclusions: Task, Activity, and Hybrid Reliabilities

In summary, NCOs vs. Officers and FORSCOM vs. TRADOC do not consistently show meaningful differences, either in terms of the agreement between these groups or in terms of the agreement among raters within the groups. Based on psychometric results, there is no advantage of one group over the other. The Task Questionnaire had the highest single-rater reliabilities of the three questionnaires, and the difficulty rating had the lowest reliability of any scale.

Validity Ratings and Rankings

We carried out several types of reliability analyses for the Attribute Validity Questionnaires, each designed to provide an estimate of the reliability (actually, inter- rater agreement) of judgments that might possibly be used in formation of synthetic prediction equations.

Non-Interest Attributes for Five Performance Areas

The first analysis focused on the reliability of judgments of the validity of attributes for the five job performance areas (i.e., the reliability of the cell means in the matrix of attribute-by-job performance areas). For this analysis, the eight vocational interest attributes could not be used because they were only rated against two of the five job performance areas.

Reliabilities were estimated by first computing an appropriate Analysis of Variance (ANOVA) and then forming an intraclass correlation coefficient using the appropriate mean squares from the ANOVA. A three-way, completely crossed ANOVA (attribute x job area x rater, raters considered random and the other two effects considered fixed) was used. ANOVAs were run within MOS for each of the following groups: Total sample; FORSCOM NCOs and Officers; TRADOC NCOs and Officers; and Combined Command NCOs, Civilians, and Officers. Volume II, Appendix I shows the ANOVA results and various intraclass correlation coefficients for each group. Note that the attribute x job area source of variance forms the "true score" variance and the three-way interaction of attribute, job area, and rater forms the error variance for the calculation of the intraclass coefficient[1]. Table 5.12 summarizes the single-rater reliabilities.

The average of the total group reliabilities (last column in the table) is .25, with moderate variance across MOS, ranging from .18 to .29. This level of reliability is acceptable, but higher levels are probably desirable and could be obtained by increasing the number of raters. It would also be possible to obtain higher reliabilities by using only officers, since their average reliability is .36, about .19 higher than the NCOs (.17). This difference in favor of the officers holds across all MOS and both commands, and to a greater extent in FORSCOM.

All Attributes for Core Technical Proficiency

The second analysis focused on the judgments of validity of all 30 attributes for Core Technical Proficiency in the MOS. This judgment is the one most likely to be used in an operational setting (Wise, Peterson, Rosse, & Campbell, 1988). ANOVAs and intraclass coefficients were computed for the same groups within each MOS as described above. However, only a two-way, crossed ANOVA was run for this problem (attributes x raters, attributes fixed and raters random). Volume II, Appendix J contains the results of these analyses and Table 5.13 summarizes the results.

In general, the reliability coefficients for these judgments are about .04 lower than those for the judgments of 22 attributes against all five performance areas. The total group reliability averages .21 across MOS, ranging from .15 to .30. Once again, officer judgments show consistently higher values than NCO judgments (across command average of .30 versus .17) -- although three MOS did show higher

[1] Intraclass coefficient$_k$ = $MS_{\text{Desc. x Att.}}/(MS_{\text{Desc. x Att.}} + MS_{\text{Rater x Desc. x Att.}})$

where MS = mean sums of squares
Desc. = descriptor or job area
Att. = Attribute

reliabilities for NCOs in TRADOC (only one difference, for 94B, was more than a trivial amount).

Eight Vocational Interest Attributes for Core Technical and General Soldiering Proficiency

The third analysis looked at the reliability of the judgments about the validity of the eight vocational interest attributes for Core Technical and General Soldiering performance areas. ANOVAs and intraclass coefficients were computed for the same groups within each MOS as for the prior two analyses. Just as for the Core Technical analyses, a two-way, crossed ANOVA was run with the same assumptions. Separate analyses were made for Core Technical and General Soldiering areas. Appendix K in Volume II contains the results of these analyses which are summarized in Table 5.14. The table shows that, for the total group, General Soldiering judgment reliabilities were very similar to those obtained for Core Technical Proficiency judgments. There was a difference of .02 between the mean reliabilities for Core Technical Proficiency and General Soldiering across the seven MOS. And as before, officers provide higher levels of reliability on average than do NCOs (with the exception of TRADOC NCOs in three MOS).

TABLE 5.12

SINGLE-RATER RELIABILITY COEFFICIENTS FOR PHASE II VALIDITY JUDGMENTS:

22 NON-INTEREST ATTRIBUTES FOR 5 JOB AREAS BY RATER SUBGROUP

		TRADOC			FORSCOM			COMBINED COMMAND			
MOS		NCO	Officer	Total	NCO	Officer	Total	NCO	Civ*	Officer	Total
16S	N	10	9	21	42	26	68	52	2	35	89
	R_{IC}	.270	.448	.324	.181	.423	.268	.201	--	.422	.282
19K	N	10	4	22	16	15	31	26	8	19	53
	R_{IC}	.259	.547	.368	.193	.283	.239	.227	.410	.334	.292
67N	N	23	8	32	11	15	26	34	1	23	58
	R_{IC}	.165	.431	.223	.239	.415	.344	.194	--	.413	.279
76Y	N	12	7	20	16	13	29	28	1	20	49
	R_{IC}	.299	.303	.307	.052	.372	.186	.153	--	.352	.233
88M	N	13	5	22	19	10	29	32	4	15	51
	R_{IC}	.251	.278	.264	.090	.449	.192	.151	.348	.389	.218
91A	N	15	11	26	15	17	32	30	0	28	58
	R_{IC}	.264	.295	.273	.136	.344	.247	.201	--	.327	.259
94B	N	8	6	14	16	12	28	24	0	18	42
	R_{IC}	.154	.202	.187	.059	.341	.175	.093	--	.306	.183

*All civilians were from TRADOC.

Total group reliabilities: mean = .25; range = .18 to .29.

TABLE 5.13

SINGLE-RATER RELIABILITY COEFFICIENTS FOR PHASE II VALIDITY JUDGMENTS:

ALL ATTRIBUTES (N=30) FOR CORE TECHNICAL PROFICIENCY (CTP) BY RATER SUBGROUP

		TRADOC			FORSCOM			COMBINED COMMAND			
MOS		NCO	Officer	Total	NCO	Officer	Total	NCO	Civ*	Officer	Total
16S	N	10	9	21	42	26	68	52	2	35	89
	R_{IC}	.209	.363	.259	.152	.307	.201	.161	--	.319	.211
19K	N	10	4	22	16	15	31	26	8	19	53
	R_{IC}	.204	.353	.266	.183	.222	.191	.198	.301	.251	.222
67N	N	23	8	32	11	15	26	34	1	23	58
	R_{IC}	.205	.543	.262	.269	.412	.356	.220	--	.452	.299
76Y	N	12	7	20	16	13	29	28	1	20	49
	R_{IC}	.285	.367	.255	.116	.368	.199	.169	--	.355	.222
88M	N	13	5	22	19	10	29	32	4	15	51
	R_{IC}	.184	.178	.179	.163	.434	.228	.159	.259	.337	.205
91A	N	15	11	26	15	17	32	30	0	28	58
	R_{IC}	.219	.214	.218	.065	.228	.125	.122	--	.209	.163
94B	N	8	6	14	16	12	28	24	0	18	42
	R_{IC}	.192	.143	.168	.109	.247	.145	.136	--	.207	.154

*All civilians were from TRADOC.

Total group reliabilities: mean = .21; range = .15 to .30.

TABLE 5.14

SINGLE-RATER RELIABILITY COEFFICIENTS FOR PHASE II VALIDITY JUDGMENTS: INTEREST ATTRIBUTES (N = 8)

FOR CORE TECHNICAL PROFICIENCY (CTP) AND GENERAL SOLDIERING PROFICIENCY (GSP) BY RATER SUBGROUP

		TRADOC			FORSCOM			COMBINED COMMAND			
MOS		NCO	Officer	Total	NCO	Officer	Total	NCO	Civ*	Officer	Total
16S	N	10	9	21	42	26	68	52	2	35	89
	R_{IC}: CTP	.120	.262	.172	.125	.385	.203	.130	--	.350	.194
	R_{IC}: GSP	.196	.436	.270	.172	.394	.251	.186	—	.414	.260
19R	N	10	4	22	16	15	31	26	8	19	53
	R_{IC}: CTP	.250	.436	.311	.230	.329	.268	.235	.335	.346	.283
	R_{IC}: GSP	.136	.219	.226	.172	.276	.226	.166	.323	.261	.222
67N	N	23	8	32	11	15	26	34	1	23	58
	R_{IC}: CTP	.353	.625	.350	.293	.480	.408	.334	—	.514	.376
	R_{IC}: GSP	.360	.383	.326	.225	.468	.370	.326	—	.424	.350
76T	N	12	7	20	16	13	29	28	1	20	49
	R_{IC}: CTP	.448	.334	.291	.172	.391	.254	.256	--	.358	.266
	R_{IC}: GSP	.398	.404	.324	.184	.258	.222	.266	—	.285	.262
88M	N	13	5	22	19	10	29	32	4	15	51
	R_{IC}: CTP	.239	.383	.270	.171	.399	.300	.194	.320	.485	.284
	R_{IC}: GSP	.244	.511	.253	.159	.393	.292	.190	.097	.521	.274
91A	N	15	11	26	15	17	32	30	0	28	58
	R_{IC}: CTP	.186	.165	.182	.052	.137	.091	.097	--	.139	.120
	R_{IC}: GSP	.194	.262	.207	.147	.301	.222	.172	--	.263	.218
94B	N	8	6	14	16	12	28	24	0	18	42
	R_{IC}: CTP	.208	.082	.096	.120	.287	.149	.144	--	.142	.136
	R_{IC}: GSP	.411	.091	.219	.151	.309	.270	.203	--	.324	.256

*All civilians were from TRADOC.

Total group reliabilities for Core Technical: mean = .24; range = .12 to .38.
Total group reliabilities for General Soldiering: mean = .26; range = .22 to .35.

TABLE 5.15

SINGLE-RATER RELIABILITY COEFFICIENTS FOR PHASE II VALIDITY JUDGMENTS:

ATTRIBUTE VALIDITY RANKINGS BY RATER SUBGROUP

		TRADOC			FORSCOM			COMBINED COMMAND			
MOS		NCO	Officer	Total	NCO	Officer	Total	NCO	Civ*	Officer	Total
16S	N	10	9	21	42	26	68	52	2	35	89
	R_{IC}	.488	.214	.303	.416	.415	.411	.423	--	.366	.380
19K	N	10	4	22	15	15	30	25	8	19	52
	R_{IC}	.435	.472	.446	.451	.364	.379	.457	.509	.381	.402
67N	N	22	8	30	11	15	26	33	0	23	56
	R_{IC}	.465	.649	.509	.484	.583	.542	.476	--	.605	.525
76Y	N	12	7	20	16	13	29	28	1	20	49
	R_{IC}	.503	.543	.490	.405	.601	.473	.442	--	.582	.480
88M	N	13	5	23	17	10	27	30	4	15	50**
	R_{IC}	.407	.330	.395	.408	.430	.378	.403	.521	.398	.384
91A	N	15	11	26	15	17	32	30	0	28	58
	R_{IC}	.541	.376	.471	.335	.397	.360	.433	--	.392	.409
94B	N	7	6	13	17	12	29	24	0	18	42
	R_{IC}	.412	.329	.353	.372	.412	.381	.385	--	.375	.372

*All civilians were from TRADOC.

**One rater's status (NCO, Officer, Civ) code was missing, but his data were included for the total group.

Total group reliabilities: mean = .42; range = .37 to .52.

Attribute Rankings

The final reliability analysis concerned the ranking of the 30 attributes in terms of validity for overall job performance. This judgment would perhaps be most useful for selecting a limited set of attributes in an experimental or operational selection battery. Once again, ANOVAs and intraclass coefficients were computed for the same groups within each MOS as described above. Again, a two-way, crossed ANOVA was run with the same assumptions as for the other two-way problems. Volume II, Appendix L contains the results of these analyses, summarized in Table 5.15.

The reliabilities for the ranking data are moderately high, and are universally higher than the reliabilities for the rating data. The lowest value in the table is .21, and it is the only value below .30. The average reliability for total group is .42 with a range across MOS of .37 to .52. The NCO and officer values show an overall difference of about .01 (favoring the officers). The TRADOC NCO reliabilities are generally higher than the TRADOC officer values, while the reverse is true for FORSCOM. It appears that Army NCOs and officers can reach high levels of agreement when asked to rank order the 30 attributes in terms of their validity for predicting overall job performance.

Summary and Conclusions: Validity Rating and Ranking Reliabilities

Certainly, the reliability of the ranking judgments is highest among the validity judgments analyzed here. However, it should be kept in mind that the rankings were completed after the judgments of the validity of the attributes for the five performance constructs. It is possible that the rankings would be much less reliable if they were completed separately, without this preamble that no doubt served to familiarize the soldiers with the attributes and to focus their attention on the several aspects of successful soldier performance. Still, it is possible that the rankings could be completed with high reliability without the arduous, comprehensive rating task. If so, t^his method would seem to be a very quick way to select the "best" predictors from the set of predictors that are presently available to the Army (if soldier judgment is to be the only form of data that will be collected).

The judgments of the validity of the 30 attributes for Core Technical Proficiency, while not so reliable as the rankings, still reach acceptable levels, especially if officers are used. Again, we note that this judgment was made in the context of judging the attributes against all five performance constructs, and the reliability of the judgment in isolation remains unknown. We think that the level of reliability would decrease if it were made in isolation, primarily because the clarity caused by contrasting Core Technical against the other four areas would be lost, at least in part. In essence, the reliability analyses neither strengthen nor weaken the case for these judgments.

Similar comments apply to the judgments of the validity of the 22 non-interest attributes for the 5 performance areas (which, of course, contain most of the Core Technical judgments) and to the judgments of the 8 vocational interest attributes for Core Technical and General Soldiering areas. These reliability values are slightly higher than those for all 30 attributes against the Core Technical judgments, and are thus acceptable, but not so high as the ranking judgments.

In sum, all the judgments about the validity of the set of attributes seem to be acceptably reliable, with rankings clearly more reliable than the other judgments; officers generally provide higher levels of reliability than NCOs; and there appear to be no meaningful differences in reliability across MOS or command.

Task, Activity, and Hybrid Questionnaires: Comparison of Instruments and Scales

In order to better understand the similarities and differences between the Task, Activity, and Hybrid Questionnaires and the ratings scales -- Frequency (FRE), Core Technical Importance (CTI), General Soldiering Importance (GSI), Overall Job Importance (OJI), and difficulty (DIF, for Hybrid only) -- several analyses were performed. We used a multitrait, multimethod correlation matrix approach to explore the ability of the various questionnaire-scale combinations to differentiate between MOS. Used in this manner, MOS are equivalent to traits, and rating scales are treated as methods, although there is no expectation that they necessarily converge. Complete matrices are presented in Volume II, Appendix M. To further investigate scale redundancy, mean intraindividual correlations between scale profiles were also produced for each questionnaire.

Differentiation Between MOS

"Discriminant validities" are correlations of item means between different MOS on the same rating scale and can be found in the monomethod-heterotrait triangles. These correlations should be low indicating that the description items are capturing differences among the MOS. FRE, CTI, and OJI ratings are expected to show the greatest discrimination. GSI ratings may be similar across MOS.

Table 5.16 presents mean discriminant validity correlations for the five scales on the three questionnaires. With mean correlations of .80 to .86 across the three questionnaires, MOS are similar on the GSI ratings. As expected, CTI shows the greatest differentiation of MOS within each questionnaire (r = .46 to .58).

TABLE 5.16

MEAN DISCRIMINANT VALIDITY
(Same Scale, Different MOS) CORRELATIONS

	Questionnaire		
Scale	Task	Activity	Hybrid
FRE	.63	.54	.50
CTI	.58	.47	.46
GSI	.86	.80	.81
OJI	.75	.63	.65
DIF	-	-	.57

The Activity and Hybrid Questionnaires appear to differentiate between MOS more than the Task Questionnaire. Differences may be due to the Task Questionnaire having more non-relevant (near zero) items that overlap between MOS. That is, relatively few tasks describe the distinguishing features of any two MOS, and the remaining tasks have the same rated importance (zero or near zero). The more shared zero-rated tasks, the greater the correlation between MOS. The Activity and Hybrid Questionnaires have fewer items, and these items tend to have fewer zero or near zero mean ratings. The difference in MOS discrimination may be more apparent than real. Predictor attributes will be picked for MOS based on the set of tasks on which the MOS differ. Thus, on the attributes selected for the prediction composite, the two MOS may be quite different. It is more important how the questionnaires lead to discrimination on selected attributes rather than how the questionnaires themselves differentiate.

Scale Redundancy

"Convergent validities" are the correlations across item means of the different rating scales within each MOS. They are presented in Table 5.17. The correlations of FRE with CTI, FRE with OJI, CTI with OJI, and GSI with OJI are all very high (averaging .95 or greater). This holds for all three questionnaires. Thus, FRE, CTI, and OJI are converging methods of describing MOS. Correlations of FRE with GSI and CTI with GSI also average near .90. On the Hybrid Questionnaire, difficulty shows similarly high correlations with the other scales with average correlations around .90.

Thus, judging from the mean profiles, one may conclude that the rating scales are providing redundant information.

TABLE 5.17

MEAN CONVERGENT VALIDITY (Different Scale, Same MOS) CORRELATIONS

	Questionnaire		
Scale	Task	Activity	Hybrid
FRE-CTI	.99	.97	.98
FRE-GSI	.91	.90	.87
FRE-OJI	.97	.97	.95
FRE-DIF	-	-	.88
CTI-GSI	.89	.86	.87
CTI-OJI	.96	.95	.96
CTI-DIF	-	-	.92
GSI-OJI	.98	.96	.97
GSI-DIF	-	-	.88
OJI-DIF	-	-	.93

The FRE and OJI scales are essentially redundant with CTI. The GSI scale shows some slight difference from the other scales but also shows the least discrimination among MOS. The DIF scale's discriminant validity correlations are low, but so is its reliability. Its discriminant information is also redundant with the other scales.

Table 5.18 presents mean off-diagonal correlations or correlations between mean profiles from different MOS and different scales. Ordinarily, one would not expect the off-diagonal correlations to be very high. However, they are as high as the discriminant validity correlations in Table 5.16, again indicate of the redundancy in the rating scales and the similarity of the MOS.

In addition to correlation among mean rating profiles between the MOS, correlation between rating scales for each instrument were also computed for each rater. Means of these intraindividual correlations are presented in Tables 5.19, 5.20, and 5.21.

TABLE 5.18

MEAN OFF-DIAGONAL (Different Scale, Different MOS) CORRELATIONS

	Questionnaire		
Scale	Task	Activity	Hybrid
FRE-CTI	.60	.49	.47
FRE-GSI	.70	.63	.59
FRE-OJI	.68	.57	.55
FRE-DIF	-	-	.46
CTI-GSI	.68	.58	.59
CTI-OJI	.65	.53	.54
CTı-DIF	-	-	.47
GSI-OJI	.80	.71	.72
GSI-DIF	-	-	.63
OJI-DIF	-	-	.56

These correlations were computed after excluding items on which the rater gave zero frequency ratings. The mean intraindividual correlations are expected to be lower than correlations between mean profiles for two reasons. The first reason is the general tendency for correlations to be lower at the individual level of analysis than at the group level (mean of individuals' scores) of analysis. The second is the removal of items with zero frequency and unrated importance. Although attenuated as expected, they also show the redundancy between FRE, CTI, and OJI ratings. GSI ratings again show more uniqueness than FRE, CTI, and OJI.

At this individual level, difficulty ratings on the Hybrid Questionnaire are essentially independent of ratings on the other scales. The low correlations between DIF, FRE, CTI, GSI, and OJI ratings (r = -.08 to .24) indicate either (a) that difficulty is relatively unrelated to frequency and importance or (b) the definition of difficulty was too broad. Difficulty was defined as "difficult to learn and perform." Many participants were uncomfortable with this definition. They argued that a task or activity may be difficult to learn but not difficult to perform or vice versa. Such ambiguity in the scale could explain why intraindividual correlations and interrater reliabilities were low.

Summary: Task, Activity, and Hybrid Instrument and Scale Comparisons

The Activity and the Hybrid Questionnaires provide better discrimination between MOS than the Task Questionnaire. For the Task Questionnaire, the lower discriminability it displays may be a function of the greater number of non-relevant items on it versus the other two questionnaires.

In general, the scales are quite redundant. It is interesting that at the individual level, the DIF scale seems rather sloppy, but at the group level, it relates well to the other scales. This contrast suggests that a large portion of the discriminant validity information for this and the other scales comes from the zero/non-zero distinction. That is, in differentiating MOS it may not matter so much how important or difficult a task is, but simply that it is some part of the job (i.e., has a non-zero frequency).

TABLE 5.19

TASK QUESTIONNAIRE: MEAN WITHIN-RATER CORRELATIONS BY MOS

	16S	19K	67N	76Y	88M	91A	94B
FRE - CTI	.68	.73	.64	.70	.69	.63	.63
FRE - GSI	.61	.56	.38	.30	.42	.44	.28
FRE - OJI	.65	.68	.58	.56	.61	.61	.51
CTI - GSI	.72	.64	.36	.35	.51	.37	.36
CTI - OJI	.77	.81	.64	.63	.69	.61	.57
GSI - OJI	.81	.77	.71	.64	.67	.72	.68

TABLE 5.20

JOB ACTIVITY QUESTIONNAIRE: MEAN WITHIN-RATER CORRELATIONS BY MOS

	16S	19K	67N	76Y	88M	91A	94B
FRE - CTI	.69	.70	.68	.70	.71	.63	.66
FRE - GSI	.59	.53	.42	.41	.45	.48	.48
FRE - OJI	.66	.65	.61	.62	.61	.61	.57
CTI - GSI	.70	.62	.49	.47	.52	.49	.50
CTI - OJI	.74	.79	.70	.71	.74	.69	.63
GSI - OJI	.78	.73	.72	.68	.65	.73	.76

TABLE 5.21

HYBRID QUESTIONNAIRE: MEAN WITHIN-RATER CORRELATIONS BY MOS

	16S	19K	67N	76Y	88M	91A	94B
FRE - CTI	.62	.71	.68	.61	.63	.63	.64
FRE - GSI	.50	.53	.27	.22	.27	.41	.24
FRE - OJI	.58	.63	.48	.47	.51	.58	.46
CTI - GSI	.63	.61	.25	.32	.46	.39	.33
CTI - OJI	.73	.79	.53	.60	.67	.61	.54
GSI - OJI	.77	.73	.64	.62	.63	.66	.65
FRE - DIF	-.08	.12	.18	.04	-.04	.02	.02
DIF - CTI	.08	.23	.24	.13	.04	.15	.07
DIF - GSI	.08	.14	.02	-.02	.04	.05	.08
DIF - OJI	.10	.22	.16	.08	.06	.11	.12

Validity Ratings and Rankings: MOS, Command, and Rank Comparisons

MOS Intercorrelations

The attribute mean profiles for each MOS were intercorrelated separately for each job area and for the rankings of attributes in terms of validity for overall job performance (validity ranking, for short). For comparison, Phase I MOS were included when available (Core Technical Proficiency, General Soldiering Proficiency, and Validity Ranking). Volume II, Appendix N presents the complete intercorrelation matrices. Table 5.22 presents a summary of these analyses. The mean MOS intercorrelations are high for General Soldiering, Effort and Leadership, Personal Discipline, and Physical Fitness/Military Bearing, moderate for Ranking, and comparatively low (although still moderately large) for Core Technical Proficiency. The Phase I MOS had lower intercorrelations than the Phase II MOS for Core Technical Proficiency and Ranking, which is sensible since the three MOS used in Phase I were selected to be maximally diverse.

MOS x Command Intercorrelations

Convergent and discriminant validities can be examined via the multitrait multimethod approach by treating MOS as "traits" and Command (FORSCOM or TRADOC) as "methods." Table 5.23 summarizes MOS x Command intercorrelations for validity ratings of the five job areas and for validity rankings. Appendix O in Volume II contains the complete intercorrelation matrices. Mean convergent validities are high for all job areas and for validity ranking. This indicates that FORSCOM and TRADOC participants within MOS provide similar validity ratings and rankings.

Mean discriminant coefficients are higher for General Soldiering Proficiency, Effort and Leadership, Personal Discipline, and Physical Fitness/Military Bearing than for Core Technical Proficiency (mean difference = .397) and validity ranking (mean difference = .208). This indicates that different attribute profiles are obtained for the seven MOS for Core Technical Proficiency and validity ranking, but not for the

TABLE 5.22

SUMMARY OF MEAN INTERCORRELATIONS FOR PHASE I AND PHASE II MOS

(Std. Dev. are shown in parenthesis)

VALIDITY RATINGS FOR JOB AREA

	CTP	GSP	EFL	DIS	FIT	RANKING
Phase II	.531 (.208)	.920 (.036)	.932 (.024)	.954 (.017)	.970 (.009)	.728 (.120)
Phase I	.393 (.099)	.936 (.028)				.508 (.141)
Phase II with Phase I	.570 (.202)	.903 (.032)				.702 (.164)
Overall	.540 (.201)	.913 (.035)				.701 (.151)

TABLE 5.23

SUMMARY OF MEAN CONVERGENT (Within MOS, Between Command) AND DISCRIMINANT (Between MOS, Within Command and Between MOS, Between Command) CORRELATIONS BY JOB AREA

VALIDITY RATINGS FOR JOB AREA (Std. Dev. are shown in parenthesis)

	CTP	GSP	EFL	DIS	FIT	RANKING
CONVERGENT						
Within MOS, Between Command (N = 7)	.897 (.018)	.892 (.038)	.900 (.062)	.913 (.050)	.949 (.019)	.936 (.026)
DISCRIMINANT						
Between MOS, Within Command (N = 42)	.509 (.204)	.878 (.046)	.900 (.032)	.916 (.033)	.946 (.021)	.695 (.127)
Between MOS, Between Command (N = 42)	.503 (.207)	.864 (.050)	.873 (.054)	.905 (.031)	.941 (.016)	.695 (.115)

remaining job areas. In other words, discriminant validity exists for these two categories (CTP, validity ranking), which can be considered MOS-specific, but not for the other four categories, which can be considered Army-wide.

MOS x Rank Intercorrelations

Convergent and discriminant validities can also be examined by treating MOS as traits and Rank (NCO or Officer) as methods. Table 5.24 summarizes these intercorrelations, and Appendix P in Volume II contains the complete intercorrelation matrices. Convergent validities are high for all job performance constructs and for validity ranking. Discriminant coefficients are lower for Core Technical Proficiency and validity ranking than for the other four job areas (mean differences = .405 for Core Technical and .219 for ranking). This pattern of results closely parallels the pattern obtained in the MOS x Command analyses. Within MOS, NCOs and Officers provide similar validity ratings and rankings. Discriminant validity is evident for Core Technical Proficiency and validity ranking, but not for General Soldiering Proficiency, Effort and Leadership, Personal Discipline, or Physical Fitness/Military Bearing.

TABLE 5.24

SUMMARY OF MEAN CONVERGENT (Within MOS, Between Rank) AND DISCRIMINANT (Between MOS, Within Rank and Between MOS, Between Rank) CORRELATIONS

	VALIDITY RATINGS FOR JOB AREA					(Std. Dev. are shown in parenthesis)
	CTP	GSP	EFL	DIS	FIT	RANKING
CONVERGENT						
Within MOS, Between Rank (N = 7)	.859 (.054)	.872 (.064)	.895 (.065)	.896 (.039)	.925 (.042)	.887 (.070)
DISCRIMINANT						
Between MOS, Within Rank (N = 42)	.517 (.202)	.878 (.046)	.897 (.061)	.911 (.046)	.938 (.028)	.698 (.128)
Between MOS, Between Rank (N = 42)	.464 (.209)	.848 (.059)	.866 (.052)	.891 (.039)	.928 (.020)	.654 (.112)

In sum, these results bear out the conclusions reached in Phase I: attribute profiles differ across MOS for Core Technical Proficiency and validity ranking, but are similar for the other four job performance areas. In addition, these data show the results to be consistent across rank and command.

Cluster Analyses of Validity Ratings and Rankings

The preceding analyses indicate that the mean attribute profiles for Core Technical Proficiency and validity ranking vary between MOS, although some similarities are clearly present. Cluster analyses of the Core Technical Proficiency and validity ranking mean profiles for the Phase I and Phase II MOS were conducted to aid the description and interpretation of the MOS relationships, and to provide a structure for subsequent study of these MOS.

The cluster structure for Core Technical Proficiency appears in Figure 5.5. Four clusters emerged. Cluster 1 is formed by Truck Mechanics (63B) and Helicopter Repairers (67N). Cluster 2 consists of MANPADS Crewmembers (16S), Armor Crewmembers (19K), and Motor Transport Operators (88M). Unit Supply Specialists (76Y) and Food Service Specialists (94B) form Cluster 3, and Administrative Specialists (71L) and Medical Specialists (91A) form Cluster 3. The distance between Clusters 3 and 4 is smaller than the distance between any other pair of clusters. Infantrymen (11B), the most distinct MOS, eventually joins Cluster 2.

FIGURE 5.5

CLUSTER ANALYSIS OF CTP MEAN RATINGS

```
Distance metric is 1-Pearson correlation coefficient
Ward Minimum Variance Method

TREE DIAGRAM
                                      DISTANCES
    0.000                                                              2.000
CTP67    --|
           |------------------------|                                  0.080
CTP63    --|                        |
                                    |----------------------------      0.795
CTP19    ---|                       |                                |
            |--|                    |                                  0.094
CTP88    ---|  |                    |                                |
               |--------|           |                                  0.203
CTP16    ------|        |           |                                |
                        |-----------|                                  0.483
CTP11    ---------------|                                            |
                                                                       1.772
CTP76    ------|                                                     |
               |----|                                                  0.200
CTP94    ------|    |                                                |
                    |--------------------------------------------      0.334
CTP91    -----|     |
              |-----|                                                  0.178
CTP71    -----|
```

The cluster formation for validity ranking appears in Figure 5.6. Clusters 1 and 2 from the Core Technical Proficiency cluster analysis are replicated in the ranking clusters. However, Core Technical Proficiency Clusters 3 and 4 are combined in the ranking data, with Administrative Specialists and Food Service Specialists entering the cluster first, Unit Supply Specialists entering next, and Medical Specialists entering last.

FIGURE 5.6

CLUSTER ANALYSIS OF MEAN RANKINGS

```
Distance metric is 1-Pearson correlation coefficient
Ward Minimum Variance Method

TREE DIAGRAM
                                     DISTANCES
    0.000                                                             2.000
RK76Y    ----|
             ||                                                       0.135
RK71L    ---|||
            |-|                                                       0.095
RK94B    ---| |
              |-----------------------------|                         0.171
RK91A    -----|                             |
                                            |-------------------      1.071
RK88M    --|                                |
           ||                               |                         0.070
RK19K    --||                               |
            |---|                           |                         0.103
RK16S    ---|   |                           |
                |----------------|          |                         0.236
RK11B    -------|                |          |
                                 |----------|                         0.732
RK67N    --|                     |
           |---------------------|                                    0.078
RK63B    --|
```

The cluster analyses clearly suggest distinctions between "Mechanic" jobs (Cluster 1) and what may be called "Soldier" jobs (Cluster 2). The interpretation of Clusters 3 and 4 is less straightforward. With the exception of Food Service Specialists, the MOS in these clusters probably require more paperwork or administrative duties than do the MOS in the other clusters.

Figures 5.7-5.10 present graphs of the Core Technical Proficiency mean profiles for one MOS from each of the four clusters, with Infantrymen included for comparison. For cognitive attributes, differentiation is greatest for Number Ability, Spatial Ability, and Mechanical Comprehension. A high level of differentiation is obtained for all of the psychomotor and physical attributes, while virtually no differentiation occurs with the temperament attributes. For vocational interest attributes, poor differentiation is obtained with Interests in Leadership, Artistic Activities, and Efficiency/Organization, while good differentiation occurs for the remaining five interests. Separate graphs for each cluster, with Infantrymen included for comparison, for Core Technical Proficiency and General Soldiering Proficiency appear in Volume II, Appendices Q and R, respectively.

FIGURE 5.7

CTP MEAN VALIDITY RATINGS, COGNITIVE: REPRESENTATIVE MOS FROM EACH CLUSTER

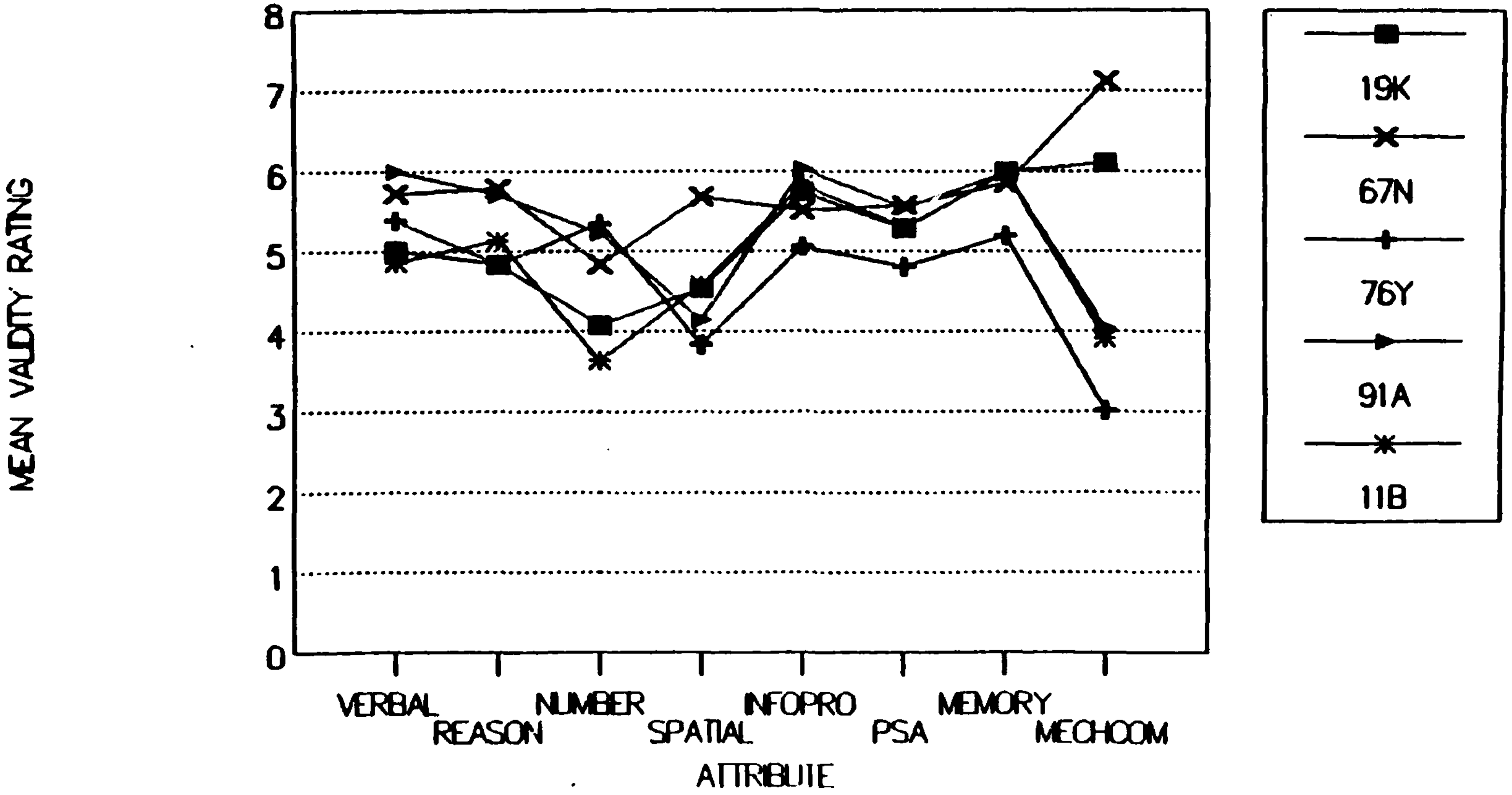


FIGURE 5.8

CTP MEAN VALIDITY, MOTOR/PHYSICAL: REPRESENTATIVE MOS FROM EACH CLUSTER

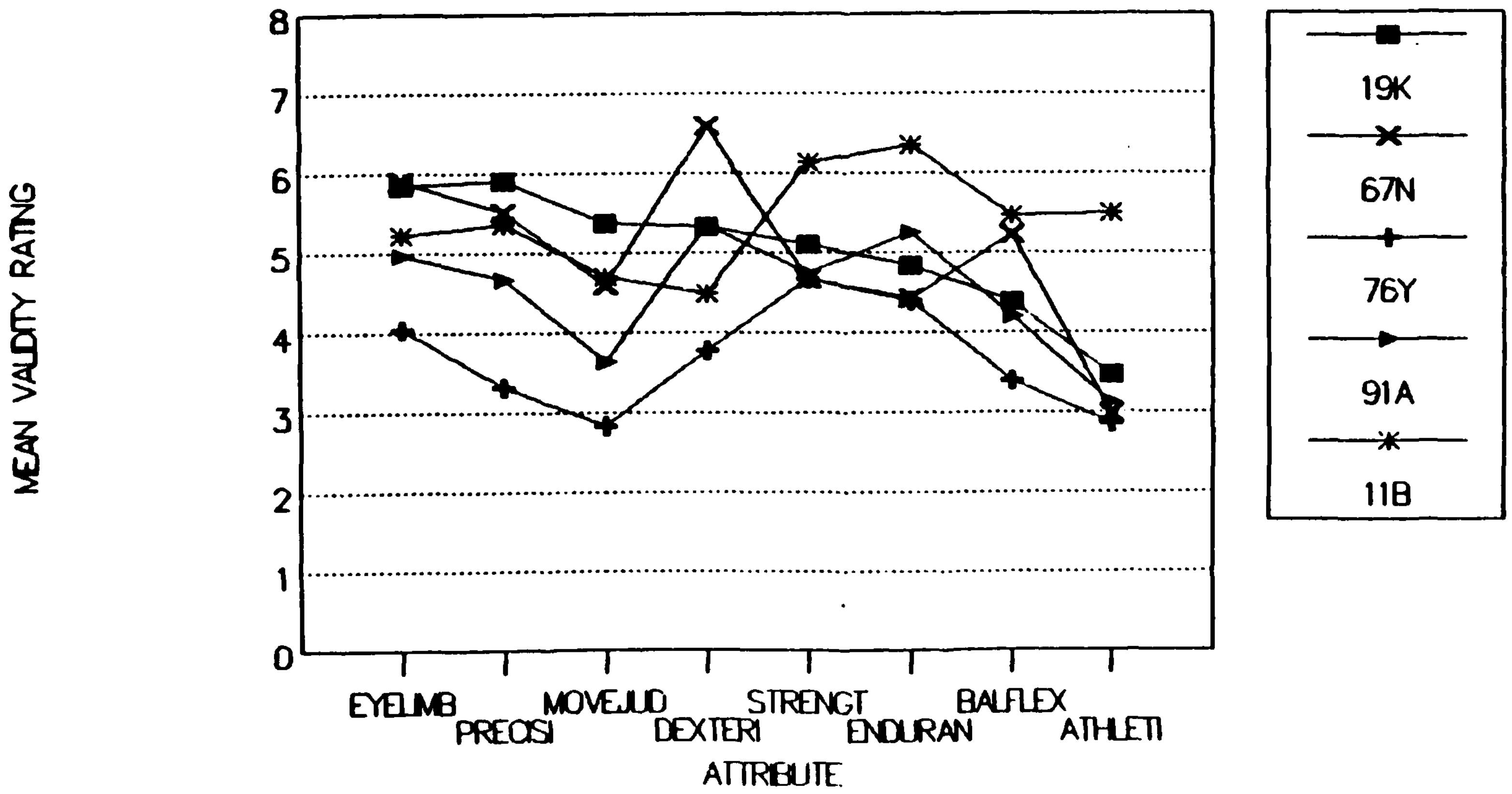


FIGURE 5.9

CTP MEAN VALIDITY RATINGS, TEMPERAMENT: REPRESENTATIVE MOS FROM EACH CLUSTER

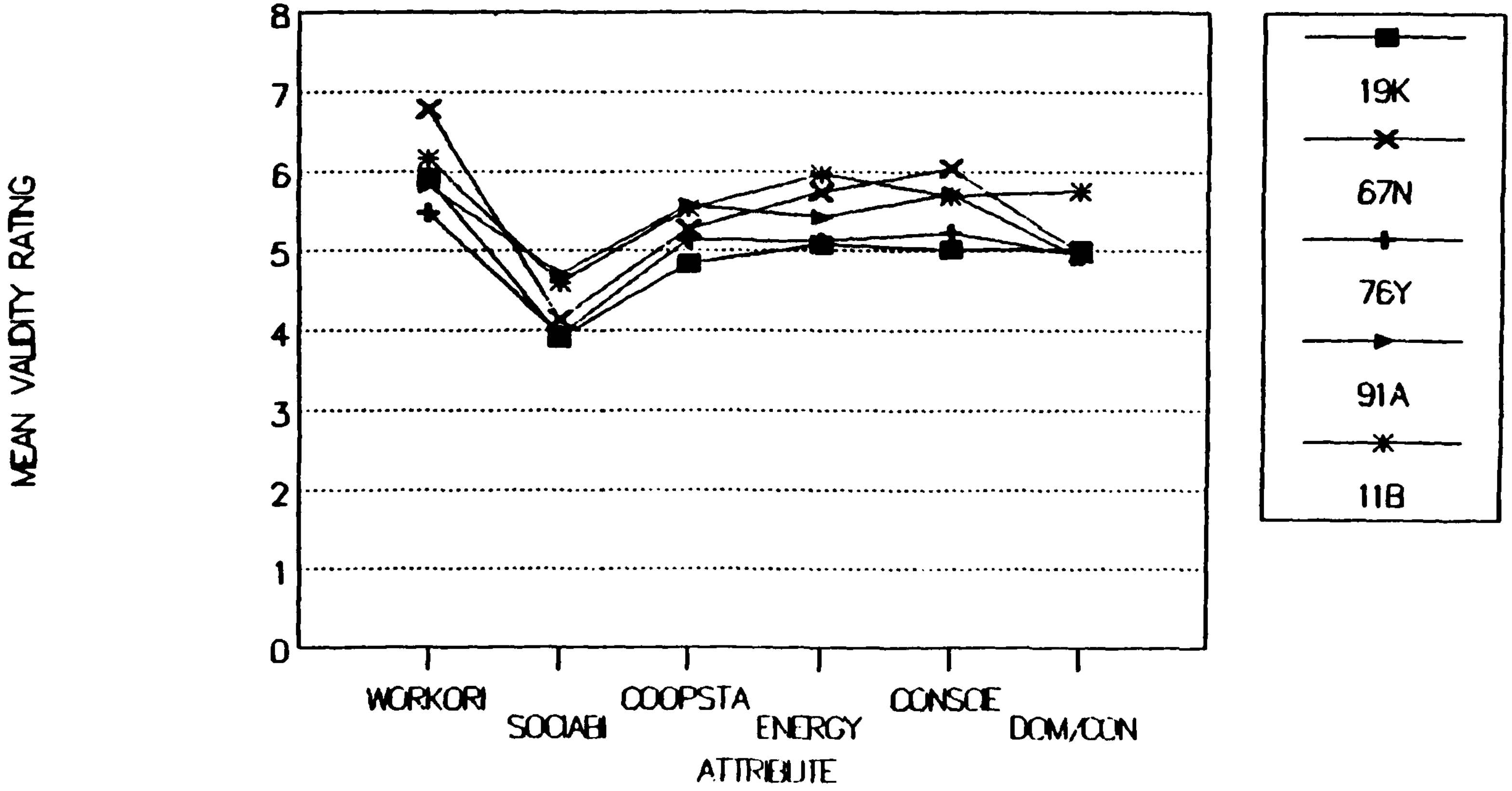
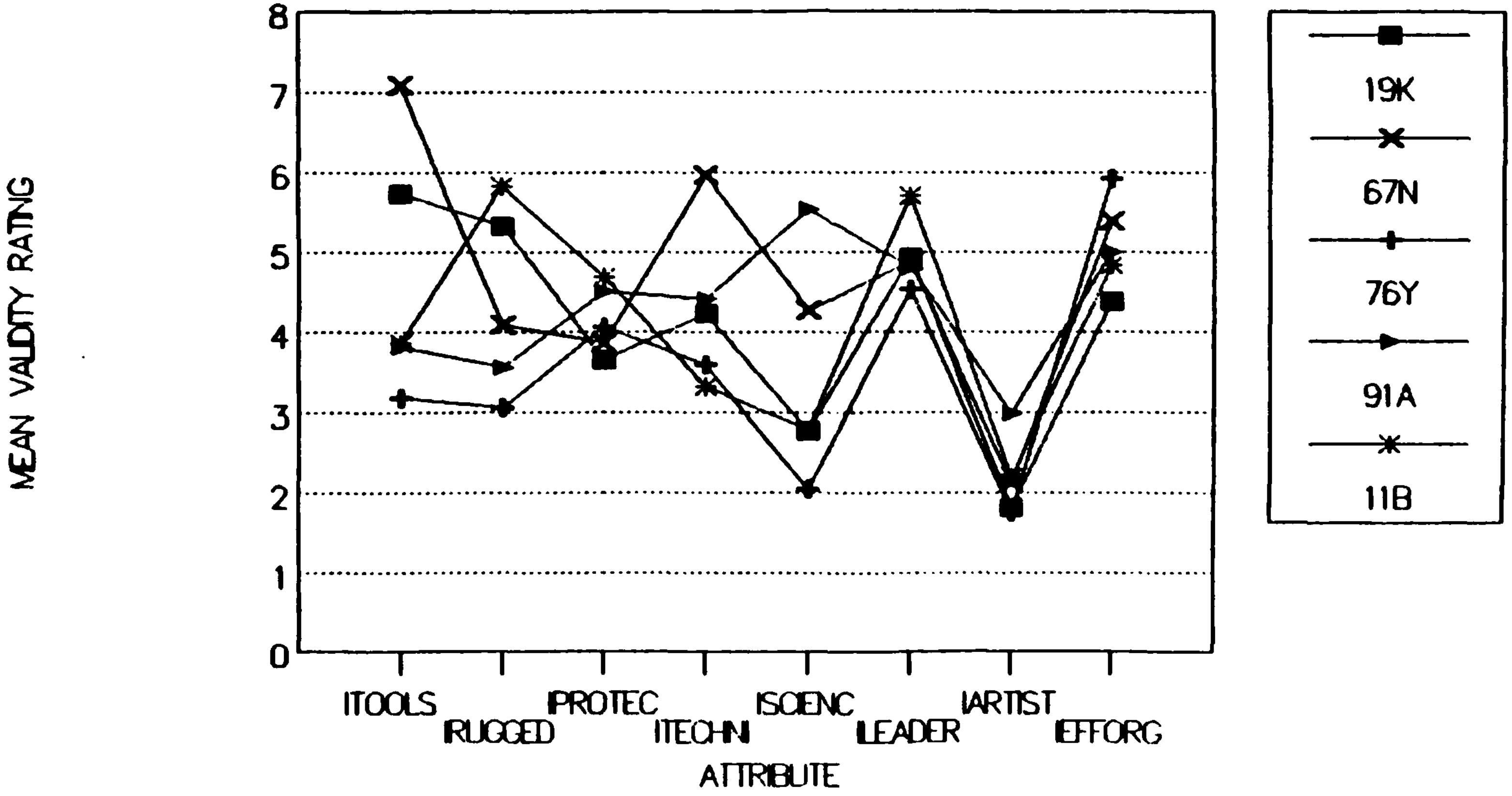


FIGURE 5.10

CTP MEAN VALIDITY RATINGS, INTERESTS:

REPRESENTATIVE MOS FROM EACH CLUSTER



Summary and Conclusions

Task, Activity, and Hybrid Questionnaires

The reliability and validity data are not necessarily definitive, but they are suggestive of several conclusions. First, we can make some conclusions concerning rater groups. The differences among the rater groups are practically nil. All rater groups agree in their assessments of their MOS, and have acceptable levels of agreement. From a psychometric standpoint, it does not seem to matter which raters are used.

Second, we can compare the rating scales. From a number of perspectives, there is a tremendous amount of redundancy in the scales. Consequently, attempts to weight job components by combining scale information is redundant. For example, weighting components by the product of frequency and Core Technical importance would be akin to simply squaring Core Technical importance.

Third, we can compare the Task, Activity, and Hybrid Questionnaires. The Task Questionnaire appears to have the edge in reliability over the other questionnaires, particularly the Hybrid model. Reliability, however, is only one aspect of the decision to recommend one questionnaire over another. The Activity Questionnaire may provide better discrimination among the MOS. However, that may be a function of the difference in the number of non-relevant items on the Task and Activity Questionnaires, and both are in very close agreement about which MOS are most similar. The way the questionnaires lead to discrimination among MOS in terms of attribute requirements is more important. The difference in their perceived coverage of MOS is negligible. Thus, the conclusions regarding preferences among questionnaires are less clear cut and need to be made in conjunction with the prediction equation analyses described in Chapter 6.

Validity Ratings and Rankings

The results found here confirm those found in Phase I. The soldiers reported that the 30 listed attributes adequately covered their jobs, although there are some suggestions for additions that could be entertained in the future.

The ratings and rankings generally reach acceptable levels of inter-rater agreement, with the rankings showing higher levels of agreement than the ratings. Officers have higher agreement levels than do NCOs, but not so much higher as to strongly favor the sole use of officers. There appears to be little difference in reliability across commands.

Mean attribute profiles for Core Technical Proficiency and validity ranking differ across MOS, but profiles for General Soldiering Proficiency, Effort and Leadership, Personal Discipline, and Physical Fitness/Military Bearing are similar across MOS. These patterns hold up across both rank of rater (NCO vs. officer) and command (FORSCOM vs. TRADOC). Cluster analyses using mean attribute profiles produce conceptually meaningful groupings of the Phase I and Phase II MOS, and examination of these profiles within attribute type across MOS show sensible, expected patterns.

This method appears to produce reliable, reasonable results. Although its use may prove not to be optional for producing synthetic prediction equations, it may prove useful for assisting specific implementation efforts, particularly when a subset of predictors must be selected.

CHAPTER 6: FORMATION OF JOB PERFORMANCE PREDICTION EQUATIONS AND EVALUATION OF THEIR VALIDITY

Scott H. Oppler (AIR), Norman G. Peterson (PDRII),and Lauress L. Wise (AIR)

Introduction

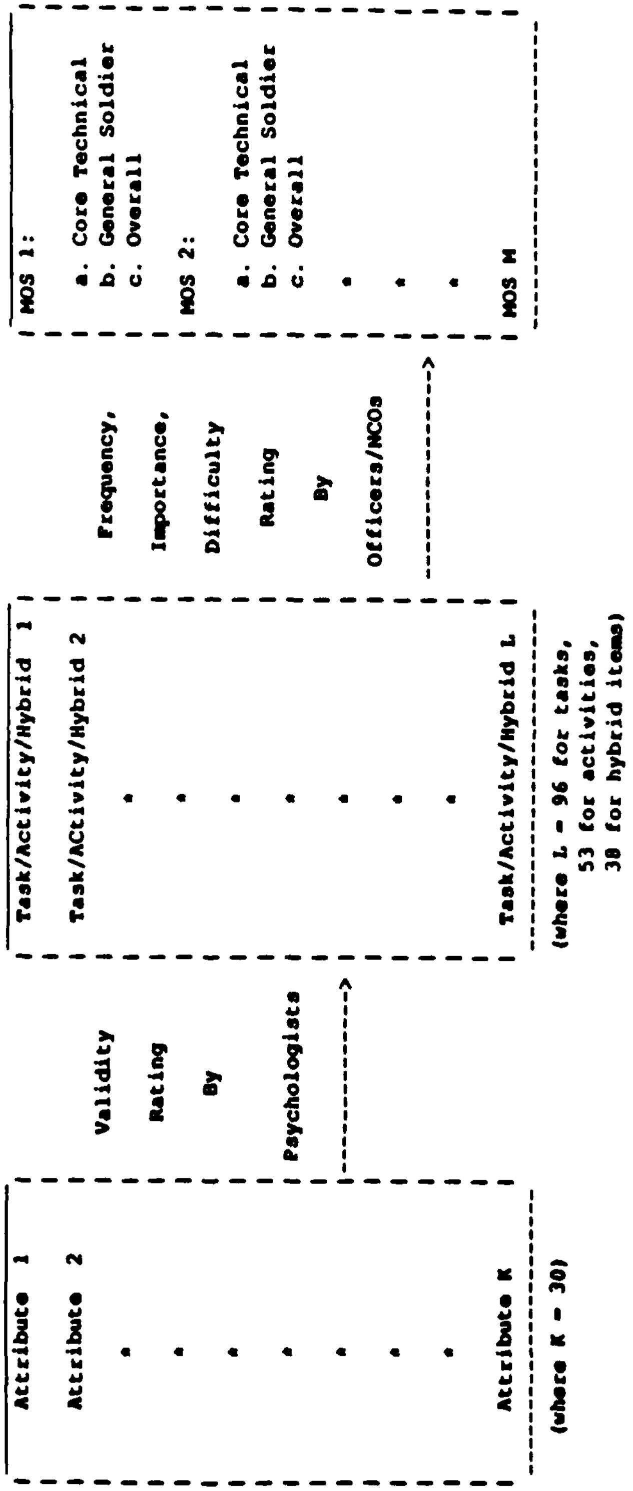
In this chapter we describe the formation of prediction equations using the ratings collected with the task, activity, hybrid, and attribute questionnaires. We also report and evaluate the results when those equations are applied to data from samples collected as part of Project A. In other words, this chapter tells what happens when we "put it all together" and attempt to predict on-the-job performance using synthetic validity methodology.

Before proceeding with the details of the methods and computations, we provide a more general overview of the elements that go into the synthetic validity methodology developed for this project. Figure 6.1 shows the elements of three of the synthetic models that we describe in this chapter. Starting at the left side of the figure, note that attribute items are tied to job descriptor components or items (task, activity, or hybrid items) by ratings of the validity of each attribute for predicting performance on each of the descriptor items. Note also that these validity ratings are made by psychologists. Thus, the attributes are here cast clearly as predictors of very discrete and relatively small pieces of Army jobs. We refer to weights obtained from these ratings as "attribute-by-component" weights.

Moving across the figure to the right, note next that the task, activity, and hybrid items are tied to a specific MOS by officers/NCOs who make ratings of the frequency, importance, or difficulty of each item with respect to a particular MOS. Note also that these ratings may be made with regard to overall performance or for slightly more specific parts of MOS job performance, such as core technical or general soldiering proficiency. Weights obtained from these kinds of ratings are referred to as "component-by-job" or "criticality" weights.

FIGURE 6.1

ELEMENTS OF "TASK," "ACTIVITY," AND "HYBRID" MODELS



In this chapter we refer to the three models depicted in this figure as the task, activity, and hybrid models because those three types of job descriptor components differ across the three models. That is, when the task questionnaire items used by officers and NCOs to describe their MOS are included in the model, then we call that model the "task" model. Note, however, that the attributes are included in all three models, tied to each of the three types of job descriptors by the psychologists' ratings of validity.

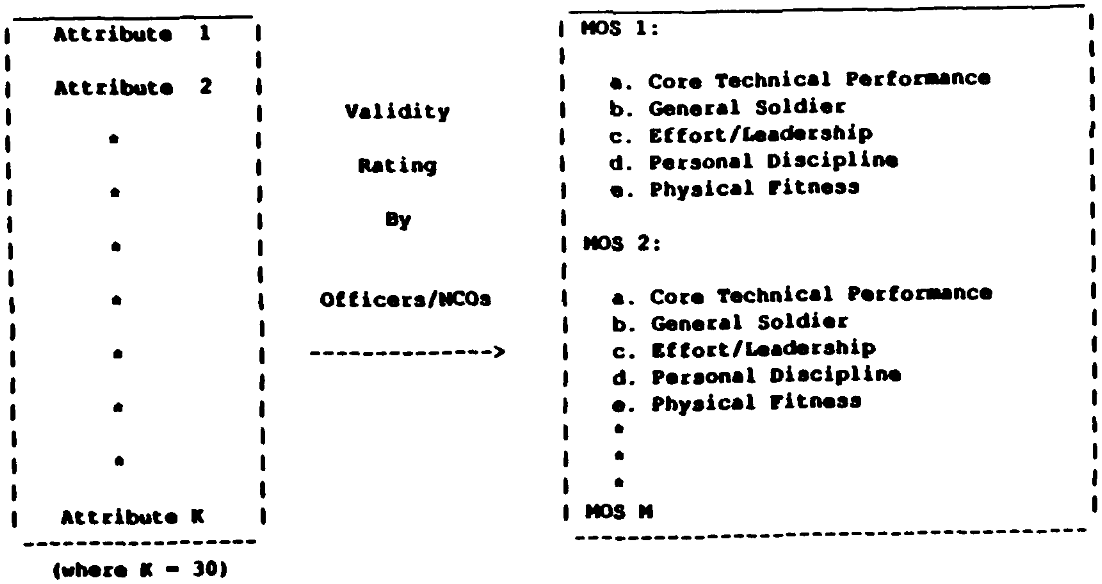
We also investigated a fourth type of synthetic validity model, and it is depicted in Figure 6.2. In essence, this model eliminates the job descriptor components and the use of psychologists for providing ratings of validities of the attributes for the job descriptor components. Instead, officers/NCOs are asked to make ratings of the validity of each of the attribute items for performance in the MOS at a fairly global level (i. e., validity for core technical performance, general soldiering performance, etc.). In this chapter, we refer to this model as the "attribute" model. Note that in this model, as compared to the other three models described above, the role of the attributes is still clearly that of a predictor, but as a predictor of a much larger piece of a job. Also, the persons providing the validity ratings are soldiers, not psychologists. The remainder of this chapter reports on two primary topics: (1) the validity of various types of synthetically formed prediction equations for the seven MOS included in Phase II of this project, and (2) the effect on the validity of synthetically formed prediction equations when subsets of psychologists (formed on the basis of familiarity with the military and relevant psychological experience) are used to provide the ratings of validity of attributes for job descriptor components.

Formation of Equations and Evaluation of Their Validity

As in the evaluations of synthetic equations derived for the three Phase I MOS (Wise, Peterson, Rosse, and Campbell, 1989), the present evaluations focus on two general criteria -- absolute and discriminant validity. Absolute validity refers to the degree to which the synthetic equations are able to predict performance in the

FIGURE 6.2

"ATTRIBUTE" MODEL ELEMENTS



specific jobs for which they were developed. For example, how well does a particular synthetic equation derived for soldiers in 19K predict core technical proficiency in that MOS? Data from Project A were used to obtain empirical estimates of these validities. The second criterion, discriminant validity, refers to the degree to which performance in each job is better predicted by the synthetic equation developed specifically for that job, than by the synthetic equations developed for the other MOS. For instance, how much better can the synthetic equation developed for 19K predict core technical performance in that MOS than the synthetic equations developed to predict core technical proficiency in each of the other MOS? Empirical estimates of correlations relevant to this criterion were also derived from data collected in Project A.

The synthetic equations whose absolute and discriminant validities are reported here were based on the four different job component models described just above. The equations based on three of these models (the activity, task, and hybrid models) required the formulation of two different sets of weights, attribute-by-component weights (for predicting MOS performance at the individual component level) and component-by-job weights (for weighting the individual component prediction equations to form an overall prediction equation). The synthetic equations based on the fourth model (the attribute model) required the specification of only one set of weights; the attribute-by-job performance construct weights (for predicting performance at more global construct levels, rather than at the component level).

We examined the degree to which the absolute and discriminant validities of the synthetic equations depend on the particular methods (described below) by which these sets of weights are respectively formulated.

Predictor Measure and Job Performance Data

The predictor measure and job performance data used in these analyses were taken from the Project A Concurrent Validation data base. The overall data set included predictor and job performance measures collected on soldiers in 19 different jobs. Seven of these jobs were the focus of the Phase II synthetic validation efforts: 16S - MANPADS Crewmember, 19K - Armor Crewman, 67N - Utility Helicopter Repairer, 76Y - Unit Supply Specialist, 88M - Motor Transport Operator, 91A - Medical Specialist, and 94B - Food Service Specialist.

The individual predictor measures included in the Project A battery have been described in detail by Peterson, Hough, Dunnette, Rosse, Houston, & Toquam, 1987. Owens-Kurtz and Peterson (1989) have described the identification of specific measures in the Project A data set corresponding to twenty-six of the thirty items in the synthetic validation project's attribute taxonomy. These twenty-six measures were used in the analyses reported here. (Thus, validity ratings were not used for the four attributes not associated with Project A measures.)

Wise, Campbell, McHenry, and Hanser (1986), and Campbell, McHenry, and Wise (1987), have described the identification and measurement of five job performance constructs of interest to the Army: job-specific proficiency (called "core technical proficiency or CTP"), general soldiering proficiency, effort and leadership, personal discipline, and physical fitness and military bearing. For the synthetic validation analyses reported here we chose to use only the job-specific proficiency measures. These measures were composed of items from written test of job knowledge and hands-on work samples. The decision to use only these measures was made for two reasons. First, and primarily, the synthetic validation project is most closely focused on the development of prediction composites for job-specific aspects of performance. Second, Wise, Campbell, and Peterson (1987) showed that the same predictor measures are optimal for a wide range of jobs in predicting all but job-specific proficiency. Significant differences across jobs were found in the predictors of job-specific proficiency. Thus, it appears that discriminant validity could not be legitimately expected for any other criterion measure.

The number of soldiers with complete data on the predictor and criterion measures in the Concurrent Validation samples corresponding to the seven Phase II MOS are reported in Table 6.1. These samples differed somewhat in terms of the heterogeneity and mean levels of the predictor scores. Also, because all were selected job incumbents, they had higher and less variable predictor scores in comparison to the overall pool from which applicants are drawn. Common practice has been to use a multivariate correction to adjust covariances and correlations for differences in heterogeneity (Lord & Novick, 1968). The 1980 Youth Population sample to which the Armed Services Vocational Aptitude Battery (ASVAB) was administered is used as the target population. This procedure corrects for effects of restriction in range due to explicit selection on the subtests of the ASVAB and incidental selection as well as due to self-selection into each occupational specialty and attrition after initial enlistment.

We used a two-step procedure to adjust for range restriction due to both sources of selection. First, we computed the covariance of the 26 predictor measures (corresponding to the attributes) for the entire Concurrent Validation sample (7,045 cases with complete predictor data) and adjusted these covariances for differences between the Concurrent Validation (CV) sample and the Youth Population in the covariances of the ASVAB subtests. This provided us with estimates of the covariances among the attribute measures for the Youth Population had all of the Project A predictor measures been administered to them. (Assumptions underlying these estimates are described in Lord and Novick, 1968).

Second, we computed covariances for each of the seven job-specific samples that included the 26 predictors plus the Core Technical Proficiency criterion construct scores. We then adjusted these covariances for differences between the job specific sample and the estimated Youth Population covariances. These corrections provided estimates of the covariances between the 26 predictors and Core Technical Performance in each of the seven MOS for the 1980 Youth Population. Table 6.2 shows the means and standard deviations of the predictor measures in the total CV sample. Tables 6.3 - 6.9 show the mean and standard deviations for each of the attribute measures in the samples for each of the seven Phase II MOS. The estimated standard deviations for the Youth Population are also shown in Table 6.2. (The means for the Youth Population are not used in the following analyses and so were not estimated.)

Method of Forming Equations

Once the covariances of the predictor and criterion measures are estimated for each job, validities for any given composite of the predictors can be estimated through relatively direct matrix manipulations. For the task, activity, and hybrid models, there are two steps in forming a synthetic predictor composite score. First, scores on individual Project A measures of the attributes are standardized, weighted (by the psychologists' ratings of validities), and summed to form a predicted score for each job component. Second, these predicted job component scores are then weighted (according to job description ratings by the soldiers/NCOs) and summed to form the predicted total job performance score. For the "attribute" model, scores on Project A measures of the attributes are weighted by the soldiers/NCOs ratings of validities for Core Technical Performance within the MOS and these products are summed to form the predicted total job performance score. We turn now to a more detailed description of the formation of these equations for the task, activity, and hybrid models.

Attribute-by-component weights. As in the Phase I analyses, three different methods were used to form the attribute-by-component weights. One method for developing prediction equations for each job component used attribute weights that were directly proportional to the attribute-by-component validities estimated by psychologists. This was called the validity method. An alternative, called the regression method, was to compute "regression" weights that took the correlations among the predictors into account. (In matrix terms, the regression weights are given by the product of the validity estimate vector with the inverse of the matrix of predictor correlations.)

TABLE 6.1
PHASE II MOS CONCURRENT VALIDATION SAMPLES WITH COMPLETE PREDICTOR AND CRITERION DATA

MOS	CV Sample
16S: MANPADS Crewmember	338
19K: Armor Crewman	394
67N: Utility Helicopter Repairer	238
76Y: Unit Supply Specialist	444
88M: Motor Transport Operator	507
91A: Medical Specialist	392
94B: Food Service Specialist	368

TABLE 6.2

PREDICTOR MEANS AND STANDARD DEVIATIONS FOR THE TOTAL CV SAMPLE

				1980
		ALL MOS		POPULATION
VARIABLE	N	MEAN	STD DEV	STD DEV
ASVAB Subtests				
GS: General Science	7045	51.40	8.13	10.00
AR: Arithmetic Reasoning	7045	52.87	7.28	10.00
VE: Verbal	7045	50.96	6.44	10.00
NO: Numeric Operations	7045	52.71	6.38	10.00
CS: Coding Speed	7045	51.28	6.68	10.00
AS: Auto/Shop Information	7045	54.14	8.53	10.00
MK: Mathematics Knowledge	7045	50.98	7.39	10.00
MC: Mechanical Comprehension	7045	53.11	8.17	10.00
EI: Electronics Information	7045	52.14	7.55	10.00
Synthetic Validity Attribute Measures				
ATTR1: Verbal Ability	7045	102.37	13.51	18.97
ATTR2: Reasoning	7045	102.44	16.46	19.27
ATTR3: Number Ability	7045	100.00	17.40	25.35
ATTR4: Spatial Ability	7045	100.00	17.43	21.18
ATTR6: Mental Info. Processing	7045	100.00	23.59	24.71
ATTR7: Perceptual Speed & Acc.	7045	100.00	17.64	20.43
ATTR8: Memory	7045	50.00	14.22	14.95
ATTR9: Mechanical Comprehension	7045	133.33	17.63	22.85
ATTR10: Eye-Limb Coordination	7045	0	14.01	14.78
ATTR11: Precision	7045	0	18.84	20.39
ATTR12: Movement Judgment	7045	6.62	9.00	9.38
ATTR13: Hand & Finger Dexterity	7045	16.73	7.76	7.86
ATTR17: Involvement in Athletics	7045	13.90	3.06	3.07
ATTR18: Work Orientation	7045	150.00	26.12	26.76
ATTR20: Cooperation/Stability	7045	150.00	26.40	26.94
ATTR21: Energy	7045	48.43	5.99	6.09
ATTR22: Conscientiousness	7045	102.48	16.52	16.66
ATTR23: Dominance/Confidence	7045	100.00	18.12	18.92
ATTR24: Interest in Using Tools	7045	200.00	32.93	34.79
ATTR25: Interest in Rugged Act.	7045	150.00	26.01	26.46
ATTR26: Interest in Protective Serv.	7045	100.00	17.03	17.20
ATTR27: Interest in Technical Act.	7045	150.00	23.55	23.57
ATTR28: Interest in Science	7045	200.00	29.23	29.51
ATTR29: Interest in Leadership	7045	40.07	8.45	8.59
ATTR30: Interest in Artistic Act.	7045	14.13	4.10	4.16
ATTR31: Interest in Efficiency & Org	7045	200.00	29.95	30.71

TABLE 6.3

PREDICTOR MEANS AND STANDARD DEVIATIONS FOR EACH MOS

16S: MANPADS Crewmember

VARIABLE	N	MEAN	STD DEV
Synthetic Validation Attribute Measures			
ATTR1: Verbal Ability	338	101.45	13.36
ATTR2: Reasoning	338	100.80	16.42
ATTR3: Number Ability	338	96.56	19.31
ATTR4: Spatial Ability	338	99.93	17.41
ATTR6: Mental Info. Processing	338	98.13	28.04
ATTR7: Perceptual Speed & Acc.	338	102.29	15.97
ATTR8: Memory	338	49.03	10.62
ATTR9: Mechanical Comprehension	338	134.30	17.25
ATTR10: Eye-Limb Coordination	338	1.50	12.83
ATTR11: Precision	338	3.30	17.73
ATTR12: Movement Judgment	338	7.72	8.14
ATTR13: Hand & Finger Dexterity	338	17.64	8.08
ATTR17: Involvement in Athletics	338	14.00	2.79
ATTR18: Work Orientation	338	148.54	26.98
ATTR20: Cooperation/Stability	338	149.82	26.67
ATTR21: Energy	338	47.95	6.14
ATTR22: Conscientiousness	338	100.10	16.78
ATTR23: Dominance/Confidence	338	101.40	18.41
ATTR24: Interest in Using Tools	338	204.51	29.84
ATTR25: Interest in Rugged Act.	338	154.80	23.84
ATTR26: Interest in Protective Serv.	338	100.53	16.74
ATTR27: Interest in Technical Act.	338	154.07	21.65
ATTR28: Interest in Science	338	202.55	27.95
ATTR29: Interest in Leadership	338	40.30	7.60
ATTR30: Interest in Artistic Act.	338	13.86	3.93
ATTR31: Interest in Efficiency & Org.	338	198.25	27.43
Performance Criterion Measure			
CTP: Core Technical Prof.	338	51.59	9.42

TABLE 6.4

PREDICTOR MEANS AND STANDARD DEVIATIONS FOR EACH MOS

19E: Armor Crewman

VARIABLE	N	MEAN	STD DEV
Synthetic Validation Attribute Measures			
ATTR1: Verbal Ability	394	104.00	14.46
ATTR2: Reasoning	394	103.92	15.83
ATTR3: Number Ability	394	100.55	18.89
ATTR4: Spatial Ability	394	102.32	16.88
ATTR6: Mental Info. Processing	394	98.44	24.50
ATTR7: Perceptual Speed & Acc.	394	103.02	17.41
ATTR8: Memory	394	50.43	10.14
ATTR9: Mechanical Comprehension	394	138.36	16.29
ATTR10: Eye-Limb Coordination	394	3.24	12.28
ATTR11: Precision	394	3.77	17.20
ATTR12: Movement Judgment	394	7.95	8.08
ATTR13: Hand & Finger Dexterity	394	16.99	7.30
ATTR17: Involvement in Athletics	394	13.87	3.04
ATTR18: Work Orientation	394	149.35	28.20
ATTR20: Cooperation/Stability	394	149.42	26.56
ATTR21: Energy	394	48.24	6.38
ATTR22: Conscientiousness	394	100.72	17.46
ATTR23: Dominance/Confidence	394	101.00	19.23
ATTR24: Interest in Using Tools	394	208.10	26.21
ATTR25: Interest in Rugged Act.	394	160.79	22.70
ATTR26: Interest in Protective Serv.	394	101.09	15.38
ATTR27: Interest in Technical Act.	394	151.71	23.55
ATTR28: Interest in Science	394	198.11	28.88
ATTR29: Interest in Leadership	394	39.31	8.50
ATTR30: Interest in Artistic Act.	394	13.68	4.03
ATTR31: Interest in Efficiency & Org.	394	198.59	28.42
Performance Criterion Measure			
CTP: Core Technical Prof.	394	102.72	15.03

TABLE 6.5

PREDICTOR MEANS AND STANDARD DEVIATIONS FOR EACH MOS

67N: Utility Helicopter Repairer

VARIABLE	N	MEAN	STD DEV
Synthetic Validation Attribute Measures			
ATTR1: Verbal Ability	238	111.24	9.78
ATTR2: Reasoning	238	111.08	11.13
ATTR3: Number Ability	238	109.16	14.51
ATTR4: Spatial Ability	238	111.73	14.81
ATTR6: Mental Info. Processing	238	102.41	13.57
ATTR7: Perceptual Speed & Acc.	238	109.18	13.10
ATTR8: Memory	238	52.06	8.26
ATTR9: Mechanical Comprehension	238	150.58	10.46
ATTR10: Eye-Limb Coordination	238	6.27	11.48
ATTR11: Precision	238	12.67	16.13
ATTR12: Movement Judgment	238	10.01	7.69
ATTR13: Hand & Finger Dexterity	238	17.53	6.73
ATTR17: Involvement in Athletics	238	14.50	2.58
ATTR18: Work Orientation	238	159.66	26.73
ATTR20: Cooperation/Stability	238	159.87	25.60
ATTR21: Energy	238	50.13	6.13
ATTR22: Conscientiousness	238	105.79	15.98
ATTR23: Dominance/Confidence	238	105.08	18.24
ATTR24: Interest in Using Tools	238	205.29	26.99
ATTR25: Interest in Rugged Act.	238	155.95	21.39
ATTR26: Interest in Protective Serv.	238	97.56	16.10
ATTR27: Interest in Technical Act.	238	148.71	23.07
ATTR28: Interest in Science	238	196.33	27.96
ATTR29: Interest in Leadership	238	39.08	8.16
ATTR30: Interest in Artistic Act.	238	13.44	4.09
ATTR31: Interest in Efficiency & Org.	238	184.53	25.80
Performance Criterion Measure			
CTP: Core Technical Prof.	238	51.01	9.28

TABLE 6.6

PREDICTOR MEANS AND STANDARD DEVIATIONS FOR EACH MOS

76Y: Unit Supply Specialist

VARIABLE	N	MEAN	STD DEV
Synthetic Validation Attribute Measures			
ATTR1: Verbal Ability	444	98.84	14.15
ATTR2: Reasoning	444	100.74	17.76
ATTR3: Number Ability	444	99.24	17.62
ATTR4: Spatial Ability	444	95.72	17.16
ATTR6: Mental Info. Processing	444	97.18	30.12
ATTR7: Perceptual Speed & Acc.	444	98.21	17.91
ATTR8: Memory	444	49.90	11.22
ATTR9: Mechanical Comprehension	444	124.00	18.18
ATTR10: Eye-Limb Coordination	444	-1.82	13.92
ATTR11: Precision	444	-4.47	19.71
ATTR12: Movement Judgment	444	4.55	11.73
ATTR13: Hand & Finger Dexterity	444	16.37	7.38
ATTR17: Involvement in Athletics	444	13.85	3.23
ATTR18: Work Orientation	444	150.79	24.35
ATTR20: Cooperation/Stability	444	150.28	24.95
ATTR21: Energy	444	48.51	5.43
ATTR22: Conscientiousness	444	104.78	15.62
ATTR23: Dominance/Confidence	444	99.10	18.10
ATTR24: Interest in Using Touls	444	190.08	32.63
ATTR25: Interest in Rugged Act.	444	140.65	26.07
ATTR26: Interest in Protective Serv.	444	95.23	17.66
ATTR27: Interest in Technical Act.	444	152.40	23.26
ATTR28: Interest in Science	444	205.66	27.51
ATTR29: Interest in Leadership	444	40.85	8.10
ATTR30: Interest in Artistic Act.	444	14.82	3.86
ATTR31: Interest in Efficiency & Org.	444	211.41	29.38
Performance Criterion Measure			
CTP: Core Technical Prof.	444	51.63	9.34

TABLE 6.7

PREDICTOR MEANS AND STANDARD DEVIATIONS FOR EACH MOS

88M: Motor Transport Operator

VARIABLE	N	MEAN	STD DEV
Synthetic Validation Attribute Measures			
ATTR1: Verbal Ability	507	96.67	12.58
ATTR2: Reasoning	507	98.90	16.84
ATTR3: Number Ability	507	92.88	17.20
ATTR4: Spatial Ability	507	97.04	16.13
ATTR6: Mental Info. Processing	507	97.36	21.81
ATTR7: Perceptual Speed & Acc.	507	96.29	18.09
ATTR8: Memory	507	49.35	10.16
ATTR9: Mechanical Comprehension	507	132.42	15.58
ATTR10: Eye-Limb Coordination	507	0.17	13.33
ATTR11: Precision	507	-1.09	18.34
ATTR12: Movement Judgment	507	6.26	8.27
ATTR13: Hand & Finger Dexterity	507	16.64	8.14
ATTR17: Involvement in Athletics	507	13.69	2.94
ATTR18: Work Orientation	507	145.43	24.67
ATTR20: Cooperation/Stability	507	145.10	26.11
ATTR21: Energy	507	47.54	5.74
ATTR22: Conscientiousness	507	100.70	15.77
ATTR23: Dominance/Confidence	507	95.85	16.81
ATTR24: Interest in Using Tools	507	211.62	29.90
ATTR25: Interest in Rugged Act.	507	149.67	23.72
ATTR26: Interest in Protective Serv.	507	100.18	17.27
ATTR27: Interest in Technical Act.	507	145.71	24.86
ATTR28: Interest in Science	507	191.02	29.36
ATTR29: Interest in Leadership	507	37.45	8.43
ATTR30: Interest in Artistic Act.	507	13.24	3.97
ATTR31: Interest in Efficiency & Org.	507	197.61	29.70
Performance Criterion Measure			
CTP: Core Technical Prof.	507	101.98	14.48

TABLE 6.8

PREDICTOR MEANS AND STANDARD DEVIATIONS FOR EACH MOS

91A: Medical Specialist

VARIABLE	N	MEAN	STD DEV
Synthetic Validation Attribute Measures			
ATTR1: Verbal Ability	392	108.73	9.64
ATTR2: Reasoning	392	106.82	13.88
ATTR3: Number Ability	392	103.29	15.36
ATTR4: Spatial Ability	392	101.42	16.50
ATTR6: Mental Info. Processing	392	101.02	15.51
ATTR7: Perceptual Speed & Acc.	392	103.46	15.87
ATTR8: Memory	392	50.59	9.30
ATTR9: Mechanical Comprehension	392	133.28	15.92
ATTR10: Eye-Limb Coordination	392	-0.89	13.67
ATTR11: Precision	392	-1.23	17.96
ATTR12: Movement Judgment	392	6.12	7.88
ATTR13: Hand & Finger Dexterity	392	15.81	7.55
ATTR17: Involvement in Athletics	392	13.69	3.30
ATTR18: Work Orientation	392	151.66	26.67
ATTR20: Cooperation/Stability	392	150.44	28.43
ATTR21: Energy	392	48.38	6.61
ATTR22: Conscientiousness	392	104.87	16.28
ATTR23: Dominance/Confidence	392	101.21	18.20
ATTR24: Interest in Using Tools	392	183.60	34.00
ATTR25: Interest in Rugged Act.	392	141.51	27.60
ATTR26: Interest in Protective Serv.	392	97.67	17.09
ATTR27: Interest in Technical Act.	392	149.12	23.70
ATTR28: Interest in Science	392	212.20	25.38
ATTR29: Interest in Leadership	392	42.28	8.23
ATTR30: Interest in Artistic Act.	392	15.83	4.17
ATTR31: Interest in Efficiency & Org.	392	196.36	28.49
Performance Criterion Measure			
CTP: Core Technical Prof.	392	102.89	15.90

TABLE 6.9

PREDICTOR MEANS AND STANDARD DEVIATIONS FOR EACH MOS

94B: Food Service Specialist

VARIABLE	N	MEAN	STD DEV
Synthetic Validation Attribute Measures			
ATTR1: Verbal Ability	368	99.93	13.00
ATTR2: Reasoning	368	98.86	18.71
ATTR3: Number Ability	368	97.40	16.99
ATTR4: Spatial Ability	368	94.80	17.71
ATTR6: Mental Info. Processing	368	96.55	23.24
ATTR7: Perceptual Speed & Acc.	368	95.46	18.27
ATTR8: Memory	368	48.81	11.15
ATTR9: Mechanical Comprehension	368	127.72	16.58
ATTR10: Eye-Limb Coordination	368	-4.39	15.12
ATTR11: Precision	368	-6.39	19.03
ATTR12: Movement Judgment	368	4.88	10.07
ATTR13: Hand & Finger Dexterity	368	14.85	8.89
ATTR17: Involvement in Athletics	368	13.40	3.06
ATTR18: Work Orientation	368	153.67	24.58
ATTR20: Cooperation/Stability	368	148.21	26.94
ATTR21: Energy	368	48.68	5.89
ATTR22: Conscientiousness	368	103.16	16.08
ATTR23: Dominance/Confidence	368	99.82	18.49
ATTR24: Interest in Using Tools	368	192.57	33.58
ATTR25: Interest in Rugged Act.	368	142.43	26.91
ATTR26: Interest in Protective Serv.	368	94.13	17.17
ATTR27: Interest in Technical Act.	368	149.04	24.78
ATTR28: Interest in Science	368	196.15	28.84
ATTR29: Interest in Leadership	368	40.11	8.55
ATTR30: Interest in Artistic Act.	368	14.82	3.89
ATTR31: Interest in Efficiency & Org.	368	219.92	30.63
Performance Criterion Measure			
CTP: Core Technical Prof.	368	52.62	9.01

A last alternative was to use zero or one weights (called the unit weight method). In this alternative, all attributes with mean validity ratings for a component less than 3.5 were given a weight of 0 and all remaining attributes were given a weight of 1.

We also investigated the effect on the validity of synthetically formed prediction equations of using subgroups of psychologists for obtaining the attribute-by-component weights. However, we defer to later in the chapter a description of that investigation.

Component-by-job or "criticality" weights. We also explored 4 different methods for forming component-by-job or "criticality" weights. The first of these methods used the mean of the officers/NCOs' (the Subject Matter Experts for Army MOS) ratings of the frequency with which the component occurred on the job as the criticality weight for each component. These ratings were on a scale from 0 to 5. Likewise, the second method assigned as the criticality weight for each component the mean of the SME ratings of the importance of these *components for core technical proficiency.* These ratings were also on a scale of 0 to 5. (Note that components which were assigned 0 ratings on the frequency scale were automatically assigned 0 ratings on the importance scale as well. See Chapter Five for a complete description of these ratings.) The third and fourth methods consisted of assigning as criticality weights multiplicative combinations of the frequency and importance mean ratings. The third method used the product of the mean ratings from SMEs on the frequency and importance scales as the criticality weight for each component, and the fourth method used the product of these mean ratings after adjusting the mean importance ratings so as to equate the variance of the frequency and importance ratings (see Kane, Kingsbury, Colton, & Estes, 1989).

Three variations for forming the four kinds of criticality weights were also explored. These variations are dubbed "threshold" methods because they assigned non-zero criticality weights to only those components with mean ratings on either the frequency or core technical importance rating scales above some specified cut-off value ("threshold")on those scales. Components with mean ratings below the cut-off value received zero weights. In the original methods, non-zero criticality weights are assigned to each of the job components (unless its computed value was actually zero).

Each variation corresponded to a different threshold level (2.5, 3.0, or 3.5). For example, the four methods that used the 2.5 threshold consisted of the following: For the frequency rating method, only components whose mean frequency ratings were 2.5 or greater were assigned those means as their criticality weights; all other components were assigned zeros. Likewise, for the core technical importance rating method, only those components receiving mean ratings of 2.5 or above on that scale were assigned non-zero criticality weights. Finally, for the two multiplicative methods, components were assigned non-zero criticality weights (corresponding to the appropriate multiplicative function) only if their mean ratings on the core technical importance rating were not less than 2.5.

"Empirical Weights." In addition to the synthetically produced predictor composites, we developed "empirical" prediction equations using least-squares regression of the 26 predictor measures against the core technical proficiency criterion composite within each of the seven MOS. The core technical proficiency criterion includes job performance measures from job knowledge tests and hands-on tests. When the same empirical data were used to estimate the validity of the empirical composites that were used to develop them, (e.g., when the equation developed on the 19K sample was applied to the 19K sample) a downward adjustment was applied to yield unbiased estimates of cross-validated coefficients for these composites (Claudy, 1978). On the other hand, no adjustments were made when we estimated the validity of the empirical equation developed for one job for predicting performance in a different job. This is because the criterion data for the other jobs were not used in the development of the empirical weights, therefore removing the possibility of positive bias due to error-fitting.

Results

Figure 6.3 shows a representation of the elements that enter into the formation of the synthetic prediction equations for the task, activity or hybrid models. Each "x" in the figure represents one equation for one MOS for one of the three models. For example, the "x" in the upper-left hand corner represents the synthetically formed equation using the mean frequency rating of the components (tasks, activities, or hybrid items) with no threshold invoked and the "validity" attribute-by-component weights. Note that there are sixteen types of component-by-job weights and three types of attribute-by-component weights, which produces 48 equations for each of the task, activity, and hybrid models for each MOS. Multiplying 48 x 3 (for the three models: task, activity, and hybrid) yields 144 equations per MOS. In addition to these 144 equations, there are three equations per MOS for the "attribute" model. These equations are formed by using the ratings of the validity of the attributes for Core Technical Performance supplied by the officer/NCO SMEs for their MOS in three ways: weights proportional to the validity estimates, regression weights, and unit weights (in the same way as explained above in **Attribute-by-component weights**).

Tables 6.10 - 6.14 contain results associated with equations based on the use of one of the sixteen component-by-job types of weights, the mean frequency ratings of activities (i.e., no threshold cut-off was used). Tables 6.10 - 6.12 show the overall weights for the synthetic equations derived for each MOS when these mean frequency weights are used in combination with each of the three types of attribute-by-component weights (i.e., validity weights, regression weights, and unit weights).

FIGURE 6.3
REPRESENTATION OF ELEMENTS INCLUDED IN SYNTHETICALLY FORMED PREDICTION EQUATIONS FOR THE "TASK," "ACTIVITY," AND "HYBRID" MODELS

Component by Job Weights

	Rating Scale:	Frequency				Importance				Freq. x Imp.				Freq. x Imp. (adj.)			
	Threshold Cut-off:	0	2.5	3.0	3.5	0	2.5	3.0	3.5	0	2.5	3.0	3.5	0	2.5	3.0	3.5
Attr. by Comp. Wgts.	Validity	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Regression	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Unit	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Note: Each "x" represents one synthetically formed equation for one MOS for one of the three models.

As explained above, composites based on the synthetic equations were correlated with Core Technical Performance only. These correlations are reported in Table 6.13. The correlations on the diagonals of the sub-matrices in this table represent the absolute validities of these composites (i.e., the correlations between each composite and CTP in the particular MOS for which it was developed), whereas the correlations on the off-diagonal represent the validities of the composites for predicting CTP in the other six MOS.

The average values of the diagonal (absolute validity coefficients) and off-diagonal correlations contained in Table 6.13 are reported in the last three columns of Table 6.14. For example, the average value of the diagonal correlations associated with the synthetic composites based on validity weights (reported in the top sub-matrix of Table 6.13) is .523 while the average off-diagonal correlation is .521. This indicates that the synthetic equations developed for the activity component model using mean frequency ratings as criticality weights and validity estimates as attribute-by-component weights are no more valid for predicting performance in the MOS for which they were intended than they are for predicting performance in the other six MOS. In other words, the average discriminant validity of these synthetic equations (which is equal to the difference between the average diagonal correlations and the average off-diagonal correlations) is approximately zero.

A similar conclusion can be made with regard to the synthetic equations based on the regression and unit attribute-by-component weights. That is, both sets of synthetic equations yield only trivial levels of discriminant validity. However, the synthetic equations associated with the three attribute-by-component weighting schemes do differentiate themselves according to absolute validity. Specifically, the synthetic equations based on the unit weights result in the highest level of absolute validity ($\underline{r}$ = .565), followed closely by those based on validity weights ($\underline{r}$ = .523). The average absolute validity for the synthetic equations based on regression weights is only .316.

The first two columns in Table 6.14 report average diagonal and off-diagonal correlations associated with the empirical prediction equations based on Project A data. (These equations, and their corresponding correlations with core technical performance in each of the seven MOS, are reported in Tables 6.15 and 6.16, respectively.) The average diagonal correlation reported in the column headed ADJEMP in Table 6.14 differs from that in the first column (labeled EMP) in that it corresponds to the average of the diagonal correlations reported in Table 6.16 after each had been adjusted for shrinkage. (As explained earlier, no such corrections were made for the off-diagonal correlations). The results in the ADJEMP column indicate that the maximum average absolute validity for this sample is .687, and the average off-diagonal correlations associated with the empirical composites is .603. These results demonstrate only marginal room for discriminant prediction across these seven MOS.

Results of analyses based on the remaining sets of synthetic equations are reported in Volume II, Appendix S. These results are summarized in Tables 6.17 - 6.19, which report findings associated with the validity, regression and unit attribute-by-component weights, respectively. Within each table, average absolute and discriminant validities are reported for synthetic equations associated with each job component model/criticality weight combination. Note that the coefficients in these tables have not been compared for statistical differences. Such comparisons will be made following the collection and analysis of data to be collected in Phase III of this research.

Inspection of these summary tables reveals several patterns in the results regarding the levels of absolute and discriminant validity associated with the different attribute-by-component weights, criticality weights, and job component models. These patterns are briefly summarized as follows.

TABLE 6.10

VALIDITY WT SYNTHETIC COMPOSITES FOR PREDICTING CTP
COMPONENT MODEL: ACTIVITY
CRITICALITY WEIGHTS: FREQUENCY

ATTRNO	VAL16S	VAL19K	VAL67N	VAL76Y	VAL88M	VAL91A	VAL94B
ATTR1	0.12	0.11	0.12	0.13	0.11	0.12	0.12
ATTR2	0.11	0.11	0.11	0.11	0.11	0.11	0.11
ATTR3	0.08	0.08	0.08	0.09	0.08	0.08	0.08
ATTR4	0.09	0.09	0.09	0.09	0.10	0.09	0.08
ATTR6	0.10	0.10	0.10	0.10	0.10	0.10	0.10
ATTR7	0.09	0.09	0.09	0.09	0.09	0.09	0.09
ATTR8	0.10	0.10	0.11	0.11	0.10	0.11	0.10
ATTR9	0.07	0.08	0.08	0.07	0.08	0.07	0.07
ATTR10	0.08	0.09	0.08	0.08	0.09	0.08	0.08
ATTR11	0.08	0.08	0.08	0.07	0.08	0.08	0.07
ATTR12	0.08	0.08	0.07	0.07	0.08	0.07	0.07
ATTR13	0.07	0.07	0.07	0.07	0.07	0.07	0.07
ATTR17	0.08	0.08	0.07	0.08	0.08	0.08	0.08
ATTR18	0.11	0.11	0.10	0.11	0.11	0.11	0.11
ATTR20	0.08	0.08	0.08	0.08	0.08	0.08	0.09
ATTR21	0.09	0.09	0.09	0.09	0.09	0.09	0.10
ATTR22	0.09	0.09	0.09	0.10	0.09	0.10	0.10
ATTR23	0.08	0.08	0.08	0.08	0.08	0.08	0.08
ATTR24	0.07	0.08	0.07	0.07	0.08	0.07	0.07
ATTR25	0.07	0.07	0.07	0.07	0.08	0.07	0.07
ATTR26	0.06	0.06	0.06	0.06	0.06	0.06	0.06
ATTR27	0.06	0.06	0.06	0.06	0.06	0.06	0.06
ATTR28	0.05	0.05	0.06	0.06	0.05	0.06	0.05
ATTR29	0.08	0.07	0.07	0.08	0.07	0.07	0.08
ATTR30	0.04	0.04	0.04	0.04	0.04	0.04	0.04
ATTR31	0.07	0.07	0.07	0.08	0.07	0.07	0.07

TABLE 6.11

REGRESSION WT SYNTHETIC COMPOSITES FOR PREDICTING CTP
COMPONENT MODEL: ACTIVITY
CRITICALITY WEIGHTS: FREQUENCY

ATTRNO	REG16S	REG19K	REG67N	REG76Y	REG88M	REG91A	REG94B
ATTR1	0.86	0.77	0.84	0.95	0.73	0.90	0.92
ATTR2	0.28	0.25	0.31	0.34	0.20	0.31	0.32
ATTR3	-0.33	-0.28	-0.30	-0.30	-0.29	-0.32	-0.31
ATTR4	0.02	-0.01	-0.06	-0.12	0.03	-0.08	-0.11
ATTR6	0.24	0.24	0.23	0.23	0.24	0.23	0.22
ATTR7	-0.02	-0.01	-0.03	-0.03	-0.00	-0.03	-0.04
ATTR8	0.18	0.17	0.19	0.19	0.16	0.19	0.19
ATTR9	-0.37	-0.28	-0.29	-0.39	-0.25	-0.36	-0.37
ATTR10	0.15	0.19	0.16	0.16	0.22	0.17	0.17
ATTR11	-0.07	-0.06	-0.06	-0.07	-0.09	-0.06	-0.05
ATTR12	0.13	0.11	0.08	0.06	0.13	0.08	0.06
ATTR13	0.16	0.19	0.19	0.20	0.19	0.19	0.19
ATTR17	0.28	0.28	0.26	0.26	0.29	0.27	0.27
ATTR18	0.15	0.19	0.18	0.18	0.21	0.18	0.18
ATTR20	0.00	-0.02	-0.01	-0.00	-0.03	-0.01	0.01
ATTR21	-0.07	-0.07	-0.08	-0.07	-0.07	-0.06	-0.05
ATTR22	0.29	0.29	0.28	0.27	0.29	0.28	0.28
ATTR23	-0.05	-0.07	-0.05	-0.05	-0.09	-0.06	-0.07
ATTR24	0.40	0.42	0.44	0.43	0.42	0.43	0.41
ATTR25	-0.07	-0.09	-0.13	-0.10	-0.09	-0.09	-0.08
ATTR26	0.12	0.12	0.12	0.13	0.11	0.12	0.12
ATTR27	0.00	0.03	0.04	0.00	0.03	0.01	-0.00
ATTR28	-0.24	-0.24	-0.23	-0.23	-0.24	-0.22	-0.25
ATTR29	0.00	-0.02	-0.02	-0.02	-0.02	-0.02	0.01
ATTR30	-0.06	-0.05	-0.07	-0.09	-0.03	-0.07	-0.09
ATTR31	0.44	0.44	0.45	0.46	0.44	0.44	0.46

TABLE 6.12

UNIT WT SYNTHETIC COMPOSITES FOR PREDICTING CTP
COMPONENT MODEL: ACTIVITY
CRITICALITY WEIGHTS: FREQUENCY

ATTRNO	UNI16S	UNI19K	UNI67N	UNI76Y	UNI88M	UNI91A	UNI94B
ATTR1	0.21	0.20	0.21	0.26	0.20	0.24	0.24
ATTR2	0.17	0.16	0.18	0.19	0.14	0.18	0.17
ATTR3	0.04	0.06	0.07	0.06	0.04	0.05	0.05
ATTR4	0.07	0.07	0.07	0.04	0.08	0.05	0.04
ATTR6	0.04	0.04	0.04	0.04	0.03	0.04	0.03
ATTR7	0.07	0.07	0.06	0.05	0.06	0.05	0.04
ATTR8	0.14	0.13	0.14	0.15	0.12	0.15	0.13
ATTR9	0.03	0.05	0.06	0.03	0.05	0.03	0.04
ATTR10	0.11	0.14	0.11	0.09	0.17	0.12	0.10
ATTR11	0.08	0.10	0.09	0.06	0.11	0.08	0.08
ATTR12	0.08	0.06	0.05	0.05	0.09	0.05	0.04
ATTR13	0.04	0.07	0.08	0.07	0.08	0.07	0.07
ATTR17	0.06	0.06	0.05	0.06	0.06	0.06	0.06
ATTR18	0.09	0.09	0.08	0.09	0.09	0.09	0.11
ATTR20	0.10	0.08	0.09	0.10	0.08	0.10	0.11
ATTR21	0.11	0.11	0.09	0.11	0.11	0.10	0.12
ATTR22	0.09	0.08	0.08	0.09	0.09	0.08	0.10
ATTR23	0.10	0.08	0.09	0.10	0.08	0.10	0.11
ATTR24	0.03	0.05	0.07	0.04	0.07	0.05	0.05
ATTR25	0.04	0.04	0.03	0.04	0.04	0.04	0.04
ATTR26	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ATTR27	0.01	0.02	0.03	0.01	0.02	0.01	0.01
ATTR28	0.00	0.00	0.00	0.00	0.00	0.01	0.00
ATTR29	0.10	0.08	0.09	0.10	0.08	0.10	0.11
ATTR30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ATTR31	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE 6.13

MULT CORRELATIONS BETWEEN SYNTHETIC COMPOSITES AND CTP
COMPONENT MODEL: ACTIVITY
CRITICALITY WEIGHTS: FREQUENCY

COMP	CTP16S	CTP19K	CTP67N	CTP76Y	CTP88M	CTP91A	CTP94B
VAL16S	0.42	0.53	0.57	0.48	0.48	0.61	0.57
VAL19K	0.42	0.53	0.57	0.47	0.48	0.61	0.56
VAL67N	0.42	0.53	0.57	0.48	0.48	0.62	0.57
VAL76Y	0.42	0.53	0.57	0.48	0.47	0.61	0.57
VAL88M	0.42	0.53	0.57	0.47	0.48	0.61	0.56
VAL91A	0.42	0.53	0.57	0.48	0.48	0.61	0.57
VAL94B	0.42	0.52	0.56	0.47	0.47	0.61	0.56
REG16S	0.27	0.33	0.30	0.26	0.28	0.42	0.33
REG19K	0.27	0.34	0.31	0.26	0.29	0.42	0.33
REG67N	0.28	0.33	0.31	0.27	0.30	0.43	0.34
REG76Y	0.27	0.31	0.28	0.26	0.27	0.41	0.33
REG88M	0.26	0.34	0.31	0.25	0.29	0.41	0.32
REG91A	0.27	0.32	0.29	0.26	0.28	0.42	0.33
REG94B	0.27	0.31	0.28	0.26	0.27	0.41	0.32
UNI16S	0.46	0.54	0.62	0.51	0.50	0.65	0.61
UNI19K	0.47	0.56	0.64	0.51	0.51	0.66	0.61
UNI67N	0.47	0.56	0.64	0.52	0.52	0.67	0.62
UNI76Y	0.47	0.55	0.62	0.53	0.50	0.66	0.63
UNI88M	0.46	0.55	0.64	0.50	0.51	0.65	0.59
UNI91A	0.46	0.55	0.62	0.51	0.50	0.66	0.62
UNI94B	0.45	0.53	0.61	0.51	0.49	0.64	0.60

TABLE 6.14

AVERAGE DIAGONAL AND OFF-DIAGONAL MULT CORRELATIONS FOR EMPIRICAL, VALIDITY, REGRESSION, AND UNIT WT COMPOSITES COMPONENT MODEL: ACTIVITY CRITICALITY WEIGHTS: FREQUENCY

MULT R	EMP	ADJEMP	VAL	REG	UNI
AVG DIAG	0.712	0.687	0.523	0.316	0.565
AVG OFF	0.603	0.603*	0.521	0.311	0.560

*Average off-diagonal multiple correlations based on empirical weights not adjusted for shrinkage.

TABLE 6.15

EMPIRICAL COMPOSITES FOR PREDICTING CTP

ATTRNO	EMP16S	EMP19K	EMP67N	EMP76Y	EMP88M	EMP91A	EMP94B
ATTR1	0.37	0.09	0.28	0.11	-0.06	0.42	0.07
ATTR2	0.41	0.17	0.06	0.12	0.38	0.23	0.24
ATTR3	0.40	0.19	-0.04	0.60	0.16	0.11	0.42
ATTR4	0.10	0.18	0.29	-0.02	-0.13	0.03	0.02
ATTR6	-0.03	0.08	-0.09	0.06	0.04	0.09	0.14
ATTR7	-0.10	0.10	-0.11	0.09	0.15	0.09	-0.08
ATTR8	-0.09	-0.09	-0.06	-0.11	0.03	-0.03	0.15
ATTR9	-0.28	0.35	0.55	0.27	0.60	0.12	0.27
ATTR10	0.02	0.08	-0.01	0.00	0.02	-0.04	-0.12
ATTR11	0.09	-0.10	0.01	-0.10	-0.08	0.02	-0.01
ATTR12	-0.01	-0.07	0.11	-0.09	-0.03	-0.01	0.05
ATTR13	0.11	0.12	0.00	-0.02	-0.00	0.03	-0.05
ATTR17	-0.07	-0.10	0.01	-0.06	-0.08	-0.10	-0.06
ATTR18	0.02	0.03	-0.10	0.22	-0.09	0.04	0.06
ATTR20	-0.05	0.04	-0.07	-0.01	0.04	0.03	-0.05
ATTR21	0.12	-0.10	0.12	0.02	0.05	-0.13	0.11
ATTR22	0.18	0.13	0.12	0.05	0.17	0.24	0.01
ATTR23	-0.17	-0.05	0.03	-0.13	-0.06	0.06	-0.03
ATTR24	-0.13	-0.01	-0.05	0.03	0.09	0.11	-0.12
ATTR25	0.23	0.25	-0.03	-0.07	0.10	0.15	-0.09
ATTR26	0.03	0.06	0.04	-0.10	0.02	0.04	0.04
ATTR27	0.12	-0.01	-0.08	-0.06	-0.03	-0.14	-0.13
ATTR28	0.04	-0.04	0.18	0.07	-0.03	0.04	0.08
ATTR29	-0.05	0.07	-0.13	0.15	0.08	-0.05	0.04
ATTR30	-0.21	0.05	0.03	-0.04	-0.01	0.07	0.10
ATTR31	-0.05	-0.04	-0.06	0.02	-0.09	-0.06	0.06

TABLE 6.16

MULT CORRELATIONS BETWEEN EMPIRICAL COMPOSITES AND CTP

COMP	CTP16S	CTP19K	CTP67N	CTP76Y	CTP88M	CTP91A	CTP94B
EMP16S	0.59	0.58	0.65	0.55	0.51	0.68	0.63
EMP19K	0.50	0.68	0.73	0.56	0.60	0.74	0.62
EMP67N	0.46	0.59	0.84	0.52	0.56	0.69	0.61
EMP76Y	0.49	0.56	0.65	0.67	0.52	0.68	0.72
EMP88M	0.47	0.64	0.73	0.55	0.64	0.72	0.62
EMP91A	0.51	0.64	0.74	0.58	0.59	0.78	0.67
EMP94B	0.47	0.54	0.65	0.62	0.51	0.67	0.78

Attribute-by-Component Weights. Similar to the findings in the Phase I analyses, the results in Tables 6.17 - 6.19 demonstrate the superiority of the unit weights over both the validity weights and the regression weights. Specifically, the absolute validities of the synthetic equations developed with the unit weights are almost always greater than or equal to the absolute validities of the corresponding equations developed with the validity weights, and they are much greater than the absolute validities of the equations developed with the regression weights. The average absolute validities across all job component model/critical weight combinations for each of these weights are reported in Table 6.20. The findings in this table indicate that the average absolute validities of the synthetic equations developed with the validity, regression, and unit weights, respectively, are .529, .314, and .554. Note that the coefficients in Table 6.20 have not been statistically compared, either.

Table 6.20 also reports the average discriminant validities of the synthetic equations developed with the three attribute-by-component weights, respectively, across all job component model/critical weight combinations. The average discriminant validities associated with the validity, regression, and unit weights are .007, .022, and .016, respectively. These findings indicate that the discriminant validities of the synthetic equations developed with the unit weights are greater that those associated with the equations developed with the validity weights, but smaller than those associated with the equations developed with the regression weights. However, as previously indicated, the absolute validities associated with the synthetic equations developed with these latter weights are so small as to preclude giving them any further consideration. Based on these findings, the remaining comments will focus on the results in Table 6.19 which summarizes results associated with the synthetic equations developed with the unit weights only.

Criticality Weights. The results in Table 6.19 provide very little evidence for differences among the four different criticality weight types (i.e., frequency weights, importance weights, and the two multiplicative combinations), although they do indicate the possible existence of differences between results associated with the "threshold" and "non-threshold" weights. Specifically, trends in the results suggest that the discriminant validities are larger for synthetic equations developed with "threshold" criticality weights than for synthetic equations developed with "non-threshold" weights. Moreover, the size of these discriminate validities varies directly with the threshold level employed. That is, the discriminant validities associated with the synthetic equations developed using the 3.5 thresholds were consistently larger than those associated with the synthetic equations developed using the 3.0 thresholds, which in turn were generally larger than the discriminant validities associated with the synthetic equations developed using the 2.5 thresholds. Note that the average discriminant validity across all synthetic equations developed using non-threshold criticality weights was .008, whereas the corresponding averages across all synthetic equations developed using the 2.5, 3.0, and 3.5 thresholds, respectively, were .012, .015, and .033.

There is a tradeoff, however, involved in obtaining the progressively larger discriminant validities associated with the increasing "threshold" weights. This tradeoff concerns absolute validity. Specifically, the equations developed with "threshold" weights tend to have lower absolute validities than the corresponding equations based on the full set of weights. Furthermore, the greater the threshold, the lower the absolute validity. Thus, the average absolute \ .idity across all synthetic equations developed using non-threshold criticality weights was .591, whereas the corresponding averages for synthetic equations developed using 2.5, 3.0, and 3.5 thresholds, respectively, were .570, .554, and .550. Based on the results in Table 6.19, it is apparent that a choice between "threshold" and "non-threshold" weights will likely depend on the relative importance placed on absolute and discriminant validity.

Component Models. The results in Table 6.19 indicate that the largest absolute validities were those associated with the synthetic equations based on the task model, followed by the hybrid, activity, and attribute models, respectively. Specifically, the average absolute validities (across criticality weights) for the synthetic equations associated with the first three of these models were .613, .554, and .532, respectively, whereas the absolute validity for the attribute model based on unit weights was .516.

On the other hand, the highest levels of discriminant validity were attained not by the task model, but by the hybrid model. In particular, the synthetic equations developed for the hybrid model with the 3.5 threshold criticality weights yielded average discriminant validities of .045. Still, for the corresponding equations developed for the task models the average discriminant validity was also good (r=.027), while yielding larger average absolute validities (r=.605 vs. r=.524). According to these results, it is apparent that choices between these two models (i.e., the task and hybrid models) will also depend on the relative importance given to each of the two types of validity.

TABLE 6.17

ABSOLUTE/DISCRIMINANT VALIDITY SUMMARY TABLE FOR VALIDITY ATTRIBUTE-BY-COMPONENT WEIGHTS

		Model			
Criticality Weights	Task	Activity	Hybrid	Attribute	Average
No Threshold					
Frequency (F)	.551/.002	.523/.002	.533/.002	---/---	.536/.002
Importance (I)	.552/.003	.525/.002	.532/.002	---/---	.536/.002
F x I	.552/.004	.521/.004	.531/.003	---/---	.535/.004
F x I/Adj	.552/.004	.521/.004	.531/.003	---/---	.535/.004
Average	.552/.003	.523/.003	.532/.003	---/---	.536/.003
2.5 Threshold					
Frequency (F)	.555/.006	.509/.004	.521/.003	---/---	.528/.004
Importance (I)	.551/.003	.517/.005	.526/.001	---/---	.531/.003
F x I	.553/.005	.516/.005	.528/.003	---/---	.532/.004
F x I/Adj	.553/.005	.516/.005	.528/.003	---/---	.532/.004
Average	.553/.005	.515/.005	.526/.003	---/---	.531/.004
3.0 Threshold					
Frequency (F)	.560/.008	.509/.004	.527/.006	---/---	.532/.006
Importance (I)	.554/.006	.514/.006	.527/.005	---/---	.532/.006
F x I	.556/.007	.514/.006	.529/.006	---/---	.533/.006
F x I/Adj	.556/.007	.514/.006	.529/.006	---/---	.533/.006
Average	.557/.007	.513/.006	.528/.006	---/---	.533/.006
3.5 Threshold					
Frequency (F)	.563/.011	.514/.009	.544/.018	---/---	.540/.013
Importance (I)	.562/.011	.524/.008	.550/.015	---/---	.545/.011
F x I	.562/.011	.524/.009	.551/.016	---/---	.546/.012
F x I/Adj	.562/.011	.524/.009	.551/.016	---/---	.546/.012
Average	.562/.011	.522/.009	.549/.016	---/---	.544/.012
Attribute X CTP	---/---	---/---	---/---	.507/.008	.507/.008
Average	.556/.007	.518/.006	.534/.007	.507/.008	.529/.007*

* Average of the column means

Table 6.18

ABSOLUTE/DISCRIMINANT VALIDITY SUMMARY TABLE FOR REGRESSION ATTRIBUTE-BY-COMPONENT WEIGHTS

	Model				
Criticality Weights	Task	Activity	Hybrid	Attribute	Average
No Threshold					
Frequency (F)	.338/.008	.316/.005	.329/.006	----/----	.328/.006
Importance (I)	.333/.007	.318/.005	.329/.005	----/----	.327/.006
F x I	.331/.014	.311/.008	.327/.011	----/----	.323/.011
F x I/Adj	.330/.012	.311/.008	.327/.011	----/----	.323/.010
Average	.333/.010	.314/.007	.328/.008	----/----	.325/.008
2.5 Threshold					
Frequency (F)	.331/.020	.292/.010	.315/.014	----/----	.313/.015
Importance (I)	.325/.014	.298/.009	.318/.008	----/----	.314/.010
F x I	.328/.019	.298/.011	.321/.013	----/----	.316/.014
F x I/Adj	.328/.018	.298/.011	.321/.013	----/----	.316/.014
Average	.328/.018	.297/.010	.319/.012	----/----	.315/.013
3.0 Threshold					
Frequency (F)	.333/.029	.288/.011	.318/.029	----/----	.313/.023
Importance (I)	.326/.021	.298/.006	.319/.024	----/----	.314/.017
F x I	.328/.025	.297/.015	.322/.027	----/----	.316/.022
F x I/Adj	.328/.025	.297/.015	.322/.027	----/----	.316/.022
Average	.329/.025	.295/.012	.320/.027	----/----	.315/.021
3.5 Threshold					
Frequency (F)	.329/.034	.281/.015	.322/.048	----/----	.311/.032
Importance (I)	.325/.031	.290/.017	.327/.049	----/----	.314/.032
F x I	.327/.033	.289/.018	.327/.049	----/----	.314/.033
F x I/Adj	.326/.032	.289/.018	.327/.049	----/----	.314/.033
Average	.327/.033	.287/.017	.326/.049	----/----	.313/.033
Attribute X CTP	----/----	----/----	----/----	.307/.030	.307/.030
Average	.329/.021	.298/.012	.323/.024	.307/.030	.314/.022*

* Average of the column means

TABLE 6.19

ABSOLUTE/DISCRIMINANT VALIDITY SUMMARY TABLE FOR UNIT ATTRIBUTE-BY-COMPONENT WEIGHTS

			Model		
Criticality Weights	**Task**	**Activity**	**Hybrid**	**Attribute**	**Average**
No Threshold					
Frequency (F)	.618/.006	.565/.005	.594/.005	---/---	.592/.005
Importance (I)	.618/.005	.569/.005	.596/.005	---/---	.594/.005
F x I	.619/.010	.556/.009	.588/.010	---/---	.588/.010
F x I/Adj	.619/.009	.556/.009	.588/.011	---/---	.588/.010
Average	.619/.008	.562/.007	.592/.008	---/---	.591/.008
2.5 Threshold					
Frequency (F)	.616/.016	.520/.013	.549/.010	---/---	.562/.013
Importance (I)	.611/.007	.533/.011	.570/.007	---/---	.571/.008
F x I	.616/.012	.531/.013	.571/.012	---/---	.573/.012
F x I/Adj	.616/.012	.531/.013	.571/.012	---/---	.573/.012
Average	.615/.012	.529/.013	.565/.010	---/---	.570/.012
3.0 Threshold					
Frequency (F)	.620/.022	.507/.011	.508/.016	---/---	.545/.016
Importance (I)	.606/.009	.520/.013	.540/.015	---/---	.555/.012
F x I	.609/.014	.521/.014	.543/.019	---/---	.558/.016
F x I/Adj	.609/.014	.521/.014	.543/.019	---/---	.558/.016
Average	.611/.015	.517/.013	.534/.017	---/---	.554/.015
3.5 Threshold					
Frequency (F)	.586/.033	.517/.020	.499/.059	---/---	.535/.037
Importance (I)	.612/.023	.522/.020	.532/.045	---/---	.555/.029
F x I	.609/.027	.522/.022	.532/.047	---/---	.554/.032
F x I/Adj	.609/.026	.522/.022	.532/.047	---/---	.554/.032
Average	.605/.027	.521/.021	.524/.050	---/---	.550/.033
Attribute X CTP	---/---	---/---	---/---	.516/.012	.516/.012
Average	.613/.016	.532/.014	.554/.021	.516/.012	.554/.016•

• Average of the column means

TABLE 6.20

COMPARISON OF ATTRIBUTE-BY-COMPONENT WEIGHTS

Attribute Weights	Average Absolute Validity	Average Discr. Validity
Validity	.529	.007
Regression	.314	.022
Unit	.554	.016

Conclusions: Validity of Synthetic Validity Models

Based on the results of Phase II analyses described above, the following conclusions seem warranted.

The use of the regression version of attribute-by-component weights should be discouraged. The somewhat larger discriminant validities associated with synthetic equations developed using these weights is not enough to compensate for the much lower absolute validities. It may be fruitful to pursue alternate approaches to the development of attribute-by-component weights. In particular, weights developed according to a "step-wise" regression approach may serve as a compromise between the unit and validity weights. That is, validity might be improved by the assignment of regression or validity attribute-by-component weights to only those attributes which significantly add to the prediction of each component. As with the unit weighting scheme, all other attributes would receive weights of zero.

Since the obtained validity coefficients do not vary consistently according to the type of ratings (frequency, importance, etc.) used to form the criticality weights, consideration should be given to discontinuing the use of one or the other of the rating scales. One possibility would be to drop the frequency ratings due to the slightly lower absolute validities of the synthetic equations in which they are employed. (Also, the face validity is probably higher for the importance ratings.)

Finally, strong consideration should be given to adopting either the task or the hybrid questionnaire as the method for obtaining component-by-job weights. Generally speaking, both the absolute and discriminant validities are higher for synthetic equations developed using task or hybrid component-by-job weights, when compared to absolute and discriminant validities for synthetic equations developed using activity component-by-job weights or the weights developed from the "attribute model" (i. e., estimates of validity of attributes for MOS made by officers/NCOs).

Validities Obtained by Using Judges Grouped According to Familiarity with the Military and Experience in Applied Personnel Psychology

During an earlier phase of the project we collected judgments from sixty-nine psychologists about the validity of thirty-one psychological attributes for predicting performance on each of 53 job activities, and each of 96 job tasks. Of the 69 experts, 46 were contract staff members and 23 were outside experts. The outside experts can be divided into five groups: members of the Scientific Advisory Committee for Project A (n=4); past-presidents of American Psychological Association (APA) Division 14 (n=5); APA Fellows (n=6); APA members (n=6); and other (n=2). (See Peterson, Rosse, & Owens-Kurtz, 1989 for a complete report of this effort.) These judgments are the data from which are derived the various attribute-by-component weights included in the analyses described above. An important research question for this project concerns the extent to which the characteristics or qualifications of these psychologists might affect the accuracy of validity of synthetic equations formed using their judgments.

The psychologists who made the validity judgments were asked about their familiarity with the military and their experiences in applied personnel, psychological activities. We also asked them about their familiarity with the Army's Project A because that project contributed much to the synthetic validation project and had received a fair amount of publicity, leading to concerns about possible contamination of judgment data due to exposure to Project A information. This information was analyzed and several subgroups of the judges were formed. These were:

Total: N=69	all judges
Low Military Familiarity: N=35	judges with less military familiarity
High Military Familiarity: N=34	judges with more military familiarity
Low Psychological Experience: N=44	judges with less applied psychological experience
High Psychological Experience: N=25	judges with more applied psychological experience
Low Military Familiarity, Low Psychological Experience: N=27	judges in the intersection of these two groups
High Military Familiarity, Low Psychological Experience: N=17	judges in the intersection of these two groups
Low Military Familiarity, High Psychological Experience: N=8	judges in the intersection of these two groups
High Military Familiarity, High Psychological Experience: N=17	judges in the intersection of these two groups
Low Project A: N=27	judges relatively unfamiliar with Project Alpha
Moderate Project A: N=16	judges familiar with Project Alpha
High Project A: N=26	judges very familiar with Project Alpha

The paper by Peterson, Rosse, & Owens-Kurtz (1989) presents full details on the formation of these subgroups. That paper also showed that there were essentially no consistent differences across these subgroups in terms of inter-rater agreement reliability.

We developed separate prediction equations using the mean validity judgments from each of these subgroups. Two types of job component models were employed, one that used mean officer/NCO ratings of the importance of 53 activities for Core Technical performance and one that used mean officer/NCO ratings of the importance of 96 task categories for Core Technical performance. The focal MOS were the seven used in Phase II.

Tables 6.21 and 6.22 show the validity coefficients obtained when scores from these two types of equations are correlated with Core Technical Proficiency criterion score, for the twelve subgroups of psychologists. The validity coefficient for the "empirical" (least squares, multiple regression, uncorrected for shrinkage) equation is also shown. Table 6.23 shows the mean validity across all seven MOS, the mean off-diagonal validity (mean validity when equations developed for one MOS are applied to other MOS), and the difference between these two means, which is an estimate of the discriminant validity of the equations.

These results show very little advantage for any of the subgroups in terms of absolute validity or discriminant validity. Most differences between absolute judge group validities are .01 and .02, the maximum difference is .03. The discriminant validity values are virtually identical across groups.

The conclusions seem to be the following. Familiarity with the military and applied psychological experience did not appreciably influence the usefulness of the validity judgments made by the psychologists used in this research. Furthermore, experience with Project A had no pronounced effect, although psychologists with moderate Project A experience provided judgments that resulted in validity coefficients that were 2 to 3 points lower than those found for psychologists with low or high Project A experience.

These are not necessarily unwelcome conclusions, since it means that the materials used in the process of collecting the attribute-by-component validity judgments did not seem to require special knowledge or experience to be used reliably or to lead to the construction of valid prediction equations using the judgments. It should be kept firmly in mind, however, that the psychologist judges were not a random sample of the general population. Virtually all had graduate level training in psychology and were (or had been) actively working in applied psychological endeavors.

TABLE 6.21

VALIDITY COEFFICIENTS FOR EQUATIONS FORMED WITH VARIOUS PSYCHOLOGIST GROUPS: MEAN IMPORTANCE RATING FOR CORL TECHNICAL ON ACTIVITY QUESTIONNAIRE AND ATTRIBUTE VALIDITY WEIGHTS

	16S	19K	67N	76Y	88M	91A	94B
Empirical*	.589	.678	.841	.672	.636	.785	.781
Total	.432	.532	.575	.484	.479	.613	.566
Low Military Familiarity	.420	.529	.570	.480	.476	.608	.560
High Military Familiarity	.427	.536	.582	.488	.483	.620	.572
Low Psychological Experience	.420	.529	.571	.480	.476	.609	.562
High Psychological Experience	.428	.538	.584	.490	.486	.621	.573
Low Military Familiarity Low Psychological Experience	.417	.525	.565	.477	.473	.604	.557
High Military Familiarity Low Psychological Experience	.426	.536	.581	.486	.482	.619	.570
Low Military Familiarity High Psychological Experience	.428	.543	.587	.489	.489	.622	.570
High Military Familiarity High Psychological Experience	.429	.537	.584	.491	.485	.621	.575
Low Project A	.426	.538	.581	.485	.485	.618	.570
Moderate Project A	.409	.515	.555	.472	.464	.595	.547
High Project A	.430	.538	.585	.490	.485	.621	.574

*Uncorrected for shrinkage

TABLE 6.22

VALIDITY COEFFICIENTS FOR EQUATIONS FORMED WITH VARIOUS PSYCHOLOGIST GROUPS: MEAN IMPORTANCE RATING FOR CORE TECHNICAL ON TASK QUESTIONNAIRE AND ATTRIBUTE VALIDITY WEIGHTS

	16S	19K	67N	76Y	88M	91A	94B
Empirical*	.589	.678	.841	.672	.636	.785	.781
Total	.446	.558	.606	.507	.502	.643	.601
Low Military Familiarity	.444	.557	.603	.504	.500	.639	.598
High Military Familiarity	.447	.560	.610	.509	.505	.646	.605
Low Psychological Experience	.442	.554	.600	.503	.497	.637	.596
High Psychological Experience	.452	.565	.616	.513	.509	.652	.610
Low Military Familiarity Low Psychological Experience	.440	.552	.595	.500	.494	.632	.591
High Military Familiarity Low Psychological Experience	.447	.559	.611	.509	.504	.646	.605
Low Military Familiarity High Psychological Experience	.458	.573	.626	.517	.515	.658	.615
High Military Familiarity High Psychological Experience	.449	.561	.611	.511	.507	.648	.608
Low Project A	.447	.562	.607	.506	.504	.645	.602
Moderate Project A	.436	.548	.592	.497	.493	.630	.589
High Project A	.452	.563	.616	.514	.506	.649	.610

*Uncorrected for shrinkage

TABLE 6.23

MEANS AND STANDARD DEVIATIONS OF DIAGONAL AND OFF-DIAGONAL VALIDITY COEFFICIENTS, AND DISCRIMINANT VALIDITIES, FOR SYNTHETIC MODEL EQUATIONS FORMED WITH VARIOUS PSYCHOLOGIST GROUPS

	Activity[1]			Task[2]		
	Diagonal	Off-Diagonal	Discriminant Validity[3]	Diagonal	Off-Diagonal	Discriminant Validity[3]
Empirical[4]	.712 (.085)	.603 (.082)	.109	.712 (.085)	.603 (.082)	.109
Total	.525 (.061)	.523 (.062)	.002	.552 (.065)	.549 (.065)	.003
Low Military Familiarity	.520 (.060)	.518 (.062)	.002	.549 (.064)	.547 (.064)	.002
High Military Familiarity	.530 (.062)	.528 (.063)	.002	.555 (.066)	.552 (.065)	.003
Low Psychological Experience	.521 (.061)	.519 (.062)	.002	.547 (.064)	.544 (.064)	.003
High Psychological Experience	.531 (.062)	.530 (.063)	.001	.560 (.066)	.557 (.065)	.003
Low Military Familiarity Low Psychological Experience	.517 (.060)	.515 (.061)	.002	.544 (.063)	.541 (.063)	.003
High Military Familiarity Low Psychological Experience	.529 (.063)	.527 (.063)	.002	.554 (.066)	.552 (.065)	.002
Low Military Familiarity High Psychological Experience	.533 (.062)	.531 (.063)	.002	.566 (.067)	.564 (.067)	.002
High Military Familiarity High Psychological Experience	.532 (.062)	.530 (.063)	.002	.556 (.066)	.554 (.065)	.002
Low Project A	.529 (.062)	.527 (.063)	.002	.553 (.065)	.551 (.065)	.002
Moderate Project A	.508 (.059)	.506 (.060)	.002	.541 (.064)	.538 (.063)	.003
High Project A	.532 (.062)	.530 (.063)	.002	.559 (.066)	.556 (.065)	.003

[1] Equation constructed using mean importance rating for Core Technical on Activity Questionnaire and attribute validity weights from various psychologist groups.
[2] Equation constructed using mean importance rating for Core Technical on Task Questionnaire and attribute validity weights from various psychologist groups.
[3] Discriminant Validity = diagonal mean minus off-diagonal mean.
[4] Uncorrected for shrinkage.

CHAPTER 7: SUMMARY AND CONCLUSIONS: SYNTHETIC VALIDATION

Norman G. Peterson (PDRII)

Research Questions Revisited

We return to the questions posed in Chapter 1 as our first means of summarizing the prior chapters.

(1) For each descriptor type, are there gaps in the "coverage" for specific MOS, as evidenced in the open-ended responses or the frequency of item endorsement?

Workshop participants expressed some concerns that not enough items appeared in the task and activity questionnaires, particularly of the "common task" type. When queried directly about the "percentage of their MOS covered", the average response ranged from 84% to 91% across the four questionnaires. No MOS was consistently viewed as being poorly "covered" by the methods. These results are found in Tables 5.1 and 5.3.

(2) What are the comparative levels of inter-judge agreement by type of item, type of judge, type of response scale?

In general, officers agree better with each other than do NCOs but NCOs are not so much lower that they are disqualified as adequate judges. There is no discernible difference between TRADOC and FORSCOM participants. Task ratings, regardless of scale, have higher agreement (.49 - .52 single-rater coefficients for frequency and importance ratings) than do activities (.31 - .36) or hybrid items (.38 - .39). The difficulty ratings of the hybrid items showed less agreement (.28). (See Tables 5.9 to 5.11 for the supporting results.) The validity ratings of the attributes showed the lowest level of agreement (.21 for ratings of all thirty attributes against Core Technical Performance). (See Tables 5.12 to 5.15 for supporting results for the attribute method).

(3) Comparatively speaking, how well do the different instruments discriminate among MOS?

The four instruments are best compared on the basis of the ratings of importance (for task, activity, and hybrid instruments) and validity (for attributes) for Core Technical aspects of the MOS. The between-MOS correlations of the mean rating profiles average .58, .47, .46, and .53 for task, activity, hybrid, and attribute methods, respectively. (See Tables 5.16 and 5.22). These comparisons are complicated by the larger number of "zero" or non-relevant items for instruments having more items (primarily the tasks, with 96 items) which tends to increase between-MOS correlations. There appears to be very little to pick between here.

(4) What response scale, or scale composite, yields the highest reliability and greatest discrimination?

The frequency and "importance for Core Technical (CTI)" scales have very similar levels of reliability across the task, activity, and hybrid questionnaires. In terms of discrimination, the CTI scale does better (about five points better, in terms of average discriminant validity coefficients between MOS). The difficulty scale on the hybrid questionnaire and the attribute validity scale both are much lower in terms of reliability, while the importance for general soldiering and overall job performance scales show no discrimination between MOS.

(5) Which judges yield the highest reliability and across-MOS discrimination?

Officers consistently show higher inter-judge agreement reliabilities than NCOs, but there do not appear to be any differences between the two types of judges in terms of discriminating between MOS. (See Tables 5.9 to 5.15).

(6) Are there any critical interactions between type of judge and type of descriptor relative to reliability or discriminability?

No.

(7) Which method of synthetic validation produces the highest estimated validity for each MOS in the Phase II sample?

On average across the seven MOS, the use of the unit weighted version of attribute-by-component weights in concert with task frequency threshold (=3.0) weights produced the highest validity coefficients (.62). (See Table 6.19).

(8) For which method(s) do the synthetically estimated validities match the Project A empirical validities most closely?

Generally speaking, the pattern of synthetically estimated validities across MOS was very similar to the pattern of Project A empirical validities, regardless of the method used to develop the synthetic composites. That is, the MOS with larger empirical validities tended to have the larger synthetic validities, and the MOS with smaller empirical validities tended to have the smaller synthetic validities. None of the methods resulted in synthetic validities exceeding the Project A empirical validities, although some were very close. (As indicated above, the use of the unit weighted version of attribute-by-task weights in concert with task frequency threshold (=3.0) weights produced average validity coefficients of .62, whereas the average empirical validity [corrected for shrinkage] was .69.) (See Table 6.19).

(9) Which method yields the maximum differential prediction?

In general, the use of the regression version of attribute-by-component weights produces the most differential prediction. (See Tables 6.17 to 6.19). However, it also produces much lower levels of absolute prediction. At acceptable levels of absolute prediction, the use of "threshold" types of component-by-MOS weights produce the best differential prediction, although some absolute validity is still lost.

(10) Which method yields the level of differential prediction that most closely matches the Project A results?

None of the methods yielded differential prediction across the seven MOS as large as the equations based on Project A empirical data (average differential validity = .08). As indicated above, the largest levels of differential prediction were associated with the regr. ssion weighted synthetic composites.

A Tentative Recommendation

We have now completed four rounds of data collection (i.e., Pretest, Pilot Test, Phases I and II) using the Task, Activity, and Attribute Questionnaires on samples that include 13 different MOS. For ten of those MOS, the instruments have been identical. In each case, it has been difficult to discern from the data analyses any incontrovertible pattern suggesting the superiority of one approach over the other. Nevertheless, there are practical constraints on how much more evidence we can try to amass to convince us of the optimum direction. Given the diversity of the MOS that have been examined, that pattern of results is not likely to change. Therefore, a recommendation is in order.

Prior to the analysis of the Phase II data, seven criteria were identified for differentiating among the four major models (task, activity, hybrid, and attribute). These criteria provide bench marks for comparing the models' abilities to function in the synthetic validity framework.

The first two criteria concern psychometric properties of the instruments themselves. First is reliability and the extent to which discernible groups of judges differ with regard to reliability. Second is content validity, or coverage of MOS by the instruments. The next two criteria refer more directly to the objectives of synthetic validation. That is, the selected job *component model* or *instrument* should lead to a set of predictor equations tailored for each MOS that (1) provide acceptable validity for each MOS, and (2) provide differential prediction among the MOS. Being the heart of the synthetic validity problem, these criteria should be afforded more weight in selecting among the competing models.

The last three criteria concern the interface between the models and the Army. Criterion five is the extent to which the models are affected by special rater requirements (other than command and rank/status). That is, are there special knowledge and/or experience requirements for raters? The sixth criterion is the potential acceptability of the model to Army policy setters. The final criterion is a catch-all. It is the collective opinion of the workshop leaders concerning ease of administration and apparent acceptability of the questionnaire to the raters.

Table 7-1 presents our ranking of the four models with respect to each of the seven criteria. Models were ranked on the first five criteria by comparing the statistical data for the models. For reliability rankings, the mean reliability for all groups across MOS on the core technical ratings were used. Percent coverage ratings were used to assess coverage. Absolute validity ranks were determined based on the cumulative ranking across the three weighting schemes on the "average" validities. Discriminant validity was treated as equal across the models. SME requirements were judged from the rater fidelity model reported in the Phase I analysis (Szenas & McHenry, 1989). In that analysis, job experience was related to rater fidelity only indirectly through job knowledge. Job knowledge was related to rater fidelity for each model, however the weights were essentially identical for each model. The hybrid model was not included in this analysis and therefore is not ranked in our comparison.

Rankings on the last two criteria are subjective. The acceptability rankings are based on our judgments about the willingness of the Army community to regard the models as credible. An edge was given to the Task Model in these rankings because, by appearance, it more closely matches the expertise of the Army subject matter experts that will be providing the ratings. The Activity and Attribute Models are further away from that expertise and require judgments outside of the typical domain of Army experience. The Hybrid model was down-graded as too short and too abstract.

Given these rankings, the Task Model appears to be the leading candidate for further use in synthetic validation. It is the number one ranking model on reliability, absolute validity, and acceptability, and it is equal to the others on discriminant validity. Coverage is the criterion on which it ranks the lowest. However, the coverage results indicate only minor problems that should have had minimal impact on overall results.

TABLE 7-1

COMPARISON OF JOB COMPONENT MODELS ON SYNTHETIC VALIDITY CRITERIA

Criteria	Ranks for:			
	Task	Activity	Hybrid	Attribute
Reliability	1	3	2	4
Coverage	3.5	2	3.5	1
Absolute Validity	1	3	2	4
Discriminant Validity	2.5	2.5	2.5	2.5
SME Requirements	2	2	--	2
Acceptability	1	2.5	4	2.5
Workshop Leader's Report	1.5	3	4	1.5
Mean Across Criteria	1.8	2.6	2.8	2.6

Note: Numbers indicates rank of each model based on comparison of statistical and subjective data. Absolute and discriminant validity rankings were double weighted in calculating mean ranks.

CHAPTER 8

ANALYSIS OF THE STANDARD SETTING DATA

Deborah L. Whetzel and Lauress L. Wise (AIR)

Description of Data

Four different standard setting instruments were administered to the Phase II sample as described in Chapter 2. Three of these instruments were designed to capture judgments about levels of performance that were considered unacceptable, marginal, acceptable, or outstanding. The fourth instrument was designed to capture judgments about how overall performance varies as a function of performance on more specific aspects of the job. In this chapter, we first present results from analyses of the data collected using the three different standard setting judgment protocols. In these analyses, we compare standards obtained from the different protocols and also compare agreement among judges on the standards they provide. It is important that there is adequate agreement among judges on the standards before the standards can be used to determine selection criteria. Finally, we turn to analyses of the data from the instrument on combining multiple standards. These analyses model the way in which judges aggregate component standards into an overall standard.

The job performance dimensions used for the standard setting exercises came from a preliminary version of the Hybrid Taxonomy (used in job analysis). The preliminary version contained a total of 24 dimensions which were reduced from job components contained in the Task Categories and Job Activities Taxonomy. Not all 24 dimensions were applicable for the Phase II jobs and thus the summary tables (e.g. Table 8.1) do not show all 24 dimensions.

For the analyses described in this chapter, we examined three proficiency categories based on the three minimum performance levels (cutoffs) that defined the performance levels described in Chapter 2. The three proficiency categories were unacceptable (less than marginal), unacceptable and marginal combined, (less than acceptable), and outstanding (greater than acceptable). The last category was described as outstanding rather than less than outstanding to enhance interpretability.

The three different standard setting protocols are referred to here as the:

- Soldier-Based Protocol (Soldier Method). Under this protocol, judges were asked to estimate the percent of current job incumbents who are performing at each of the four levels of acceptability (e.g., what percent are unacceptable) on a given performance dimension. This approach assumes that empirical data on soldier performance are available (in the form of hands-on tests scored GO/NO-GO) on a representative sample of the soldiers in question so that these "percent-performing" estimates can be related to actual performance scores.

- Critical Incident Protocol (Incident Method). Under this protocol, judges were presented with incidents that reflected varying levels of effectiveness on a particular performance dimension and asked to judge, for each incident, the acceptability level of soldiers whose typical performance was described by the incident.

- Task-Based Protocol (Task-Hypothetical Soldier, Task-Detailed Percent Go, and Task-Abbreviated Percent Go Methods). Under this protocol, judges were presented with a list of tasks within each performance dimension (possibly from different MOS) and asked to make judgments about minimum percent-go scores that a soldier should achieve to qualify as marginal, acceptable, and outstanding performers. Three types of judgments were collected. For some dimensions, the judges were presented detailed sets of hands-on test score sheets and corresponding summary percent-GO scores for 10 hypothetical soldiers and asked to rate the acceptability of each hypothetical soldier (Task-HS Method). In this condition, the judges were also asked to rate the minimum percent-GO score for each level of acceptability on each task and across tasks used to illustrate the dimension (Task-DPG Method). Under the third, abbreviated approach, judges were given a list of tasks without detailed percent-GO scores or actual score sheet examples and asked to rate minimum percent-GO scores for tests on these types of tasks (Task-APG Method).

Converting Standard Setting Results to a Common Metric

The five different standard setting methods involved judgments that used very different metrics. The Soldier-Based method asked about the percent of soldiers performing at each acceptability level; the Critical Incident method used a series of discrete behavioral items; and the Task-Based methods used judgments about acceptable levels of percent-GO scores.

A critical question in this research was the extent to which the different methods led to similar or distinct ability requirements. To answer this question, it was necessary to convert the standards derived from each approach onto a common metric. It would then be possible to determine whether one of the methods led to significantly stricter or more lenient standards than the others and also to compare the level of agreement among judges using this same metric.

We chose the Soldier-Based metric (percent of soldiers performing at each level) as the basis for comparison, to a large extent, because it was included as a check on the other two approaches. If standards set with the other methods led to very different assessments of the percent of soldiers performing at each level (in comparison to the judges direct assessment), then the validity of these methods would be questionable.

Data from Project A on samples of incumbents in each of the MOS were used to estimate the percent of soldiers performing above or below each of the standards set. A brief description of the conversion process for each type of instrument is given here.

Critical Incident Method. Each of the behavioral incidents rated by the judges had been previously assigned "level of effectiveness" scores on the basis of retranslation workshops conducted during the development of the Project A rating scales. Thes: effectiveness scores were on a nine-point scale, with one being the least effective and nine being the most effective. Subsequently, seven-point rating scales were developed using summaries of the incidents as anchors. Incidents with effectiveness levels of one to three were used to anchor the first two points on the seven-point scale. Incidents with effectiveness levels four to six were used to anchor the middle range of the seven-point scale, and incidents with effectiveness levels seven to nine were used to anchor the upper two points in the seven-point scale. We used the following translation to approximate the conversion between these two scales:

$$\text{EFF (7-point)} = .75 * \text{EFF (9-point)} + .25$$

This formula translates a score of 1 to a score of 1 on the seven-point scale, a score of 9 to a score of 7 and a score of 5 to a score of 4.

For each judge in our workshops, we examined all of the incidents rated at one level (e.g., unacceptable) and found the one with the highest effectiveness level. We examined all of the incidents rated at the next higher level (e.g., marginal) and found the one with the lowest effectiveness level. We then took the average of these two effectiveness levels as the dividing point between the two acceptability levels and computed the corresponding value on the seven-point scale. These dividing points or "cut scores," were computed for each judge and acceptability level.

Next, we looked at the empirical distribution of ratings of job incumbents on the same seven-point effectiveness scale. (We examined the average rating across all peers and supervisors so that the rating for each soldier was not necessarily an integer.) For each of the cut scores computed from the incident ratings, we determined the percent of job incumbents who had an average rating (below) the cut score based on the Project A ratings. These percents were used as the estimate of the percents of soldiers performing in the categories defined ɒy the incident above (below) the cut score.

Task-Based Methods. Two types of Task-Based ratings were obtained. The first type was a rating of ten hypothetical soldiers who were described in terms of the percent of steps in each sample Hands-on (HO) task that had been scored "GO" (correctly performed). These values had been derived from the Project A Concurrent Validation data by dividing actual job incumbents into deciles based on their total HO score and computing the average percent-GO on each task (and for all tasks on a particular dimension) separately for each decile group. For these ratings, we merely counted the number of hypothetical soldiers rated at each level and multiplied by 10. Thus, if a judge rated the first three soldiers as unacceptable, we estimated a 30 percent unacceptable rate.

The second type of task rating was an estimate of the minimum percent-GO score required to achieve a given level of performance. In order to convert these values into the "percent performing" metric, we examined the ten "hypothetical" soldiers. These were the average of the observed percent-G) scores for soldiers in each decile group. (The percent-GO) score for the first soldier was the average of the scores for all soldiers in the bottom ten percent. The second soldier's score was computed as the average of the next worst scoring tenth of the sample, and so on.) We then interpolated the performance percentile level corresponding to a particular minimum percent-GO score. For example, suppose that the average percent-GO for the second decile group was 58 and the average percent-GO score for the third decile group was 63 and that a given judge r:ported a minimum acceptable percent-GO score of 61. Since 61 is 3/5ths of the distance from 58 to 63, we would estimate a percentile score that was 3/5ths of the way between the 15th percentile (the mean for the second deci': group) and the 25th percentile (the mean for the third decile group). In this case, we would estimate that 21 percent of soldiers were performing below the acceptable level (and 79 percent at or above the acceptable level).

It should be noted that empirical data were not available for some of the task dimensions (Abbreviated Percent GO only). Therefore, no attempt was made to convert these data into the common metric. Since this was the case only for the Task-Based APG method, their exclusion increased the comparability of the APG and DPG ratings by eliminating differences in the dimensions covered.

Analyses of the Soldier Method Data

Editing steps. Before analyzing the Soldier-Based data, we checked each record (combination of judge and dimension) to see that the percents added to 100 across the four different acceptability levels. The original documents were checked for all records flagged by this edit to be sure that no data entry errors had occurred. There were 135 cases with some kind of problem, either those that did not add to 100 or were simply missing. We resolved the discrepancies by setting missing data equal to zero and by setting to missing the records that did not add to 100.

Analysis by performance dimension. Table 8.1 shows the means and standard deviations of the judges ratings of the percent of soldiers performing at each acceptability level for each combination of performance dimension and MOS. There are some distinct differences in the judges estimates of soldiers' ability across different MOS and dimensions. For example, 16S had high acceptability ratings for performance dimension 7 (Detect Targets), but relatively lower acceptability ratings on dimension 15 (Operate Vehicles). These differences reflect, in part, the appropriateness or importance of the dimension for the MOS. In the present research, the dimensions to be rated for each MOS were selected in advance of collecting job description information. Under an operational scenario, we would collect and analyze job descriptions first and then apply standard setting methods to those dimensions judged most critical (e.g., relevant, important, frequently performed, whatever).

The standard deviations in Table 8.1 are a measure of the degree of agreement among judges. These numbers also give an indication of the potential appropriateness of the dimension for the MOS. When there is more significant disagreement among judges, it may be because the dimension is poorly described or is not clearly appropriate for the MOS in question. To a certain extent, the standard deviations are related to the means--when there is more disagreement, the means tend to be closer to 50 percent of soldiers performing at a particular proficiency level. (Very high or low scores are only possible nearly all of the judges consistently give high or low ratings.) In some cases, however, the standard deviations are greater than the means (e.g., the percent of 16S rated unacceptable on Operate Vehicles or the percent of 88M rated unacceptable on Navigate). This can only happen when the distribution of ratings is highly skewed with most judges giving low ratings (hence a low mean) and a few judges giving very high ratings (leading to a large standard deviation).

Analysis by type of judge. Table 8.2 shows the mean ratings (averaged across different dimensions) for each type of judge and MOS. It is interesting to note the similarities across judge types. At all three levels of proficiency, the overall average percent of soldiers does not differ by more than three points between FORSCOM and TRADOC, and by more than five points between NCOs and officers.

Table 8.3 shows estimates of single-rater reliability for each type of judge and each acceptability level. There were significantly lower levels of reliability in the ratings of Unacceptable and Less than Acceptable levels for the TRADOC judges in comparison to the FORSCOM judges. This may be due to a greater heterogeneity among the TRADOC judges as this group frequently included civilians responsible for course development in addition to NCOs and Officers who work directly with students. In addition, the FORSCOM judges were likely to be more familiar with current performance levels in the field as opposed to in training. There were no significant differences between the reliability estimates for NCOs and Officers.

TABLE 8.1
SOLDIER METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY DIMENSION AND MOS (TOTAL PRE-DELPHI SAMPLES)

Level/	16S	19K	67N	76Y	88M	91A	94B	Avg.
Performance Dimension	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD
Percent Unacceptable								
2. Crew Served Wpns	14/18	08/07	.	.	.	.	.	12/13
3. Tactical Mvmnts	12/12	10/09	.	.	.	.	.	11/11
4. Navigate	.	.	.	.	21/22	.	.	21/22
5. First Aid	.	.	.	.	.	16/18	.	16/18
7. Detect Targets	8/07	10/09	10/14	.	.	.	.	11/11
8. Repair Mech. Sys	16/14	10/08	12/10	.	15/12	.	.	14/12
10. Use Tech Refs.	.	.	.	20/19	.	.	.	20/19
11. Pack and Load	17/19	07/10	.	19/19	14/14	.	14/10	16/15
13. Operate/Install	.	.	11/13	.	.	.	15/15	13/14
15. Operate Vehicles	21/26	05/07	.	.	8/06	.	.	14/14
16. Type	.	.	.	25/24	.	.	.	25/24
17. Record Keeping	.	.	17/15	16/17	20/19	19/19	.	18/18
18. Oral Comm.	16/14	12/12	.	.	.	13/12	.	16/14
19. Written Comm.	.	.	.	27/25	.	15/12	.	21/19
22. Medical Treatmnt						11/11		11/11
23. Food Preparation	.	.	.	.	.	.	13/12	13/12
24. Leadership	.	.	.	.	.	16/15	.	16/15
Average	15/16	09/09	13/13	21/21	16/15	21/15	14/12	16/15
Sample Size	563	378	162	235	250	342	129	1807
Percent Less Than Acceptable								
2. Crew Served Wpns	32/20	23/16	.	.	.	.	.	28/17
3. Tactical Mvmnts	33/16	27/21	.	.	.	.	.	30/19
4. Navigate	.	.	.	.	40/23	.	.	40/23
5. First Aid	.	.	.	.	.	37/26	.	37/26
7. Detect Targets	24/15	28/21	26/18	.	.	.	.	28/19
8. Repair Mech. Sys.	42/21	30/19	26/13	.	37/20	.	.	37/19
10. Use Tech Refs.	.	.	.	40/22	.	.	.	40/22
11. Pack and Load	38/25	25/23	.	42/23	33/21	.	11/07	37/23
13. Operate/Install	.	.	27/20	.	.	.	10/07	18/14
15. Operate Vehicles	36/23	18/16	.	.	24/15	.	.	34/19
16. Type	.	.	.	47/27	.	.	.	47/27
17. Record Keeping	.	.	37/22	39/24	42/23	26/11	.	30/20
18. Oral Comm.	36/20	31/24	.	.	.	36/23	.	37/23
19. Written Comm.	.	.	.	52/26	.	40/21	.	46/24
22. Medical Treatmnt.	.	.	.	.	.	30/21	.	30/21
23. Food Preparation	.	.	.	.	.	.	13/17	13/17
24. Leadership	.	.	.	.	.	43/23	.	43/23
Average	34/20	26/20	29/18	35/24	35/20	35/21	11/10	31/19
Sample Size	563	378	162	235	250	342	129	1807

Note A total of 24 performance dimensions were available for standard setting. However, not all were relevant for Phase II MOS. Only the relevant dimensions are shown in the Tables.

(Continued)

TABLE 8.1 (CONTINUED)
SOLDIER METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY DIMENSION AND MOS (TOTAL PRE-DELPHI SAMPLES)

Level/ Performance Dimension	16S Mn/SD	19K Mn/SD	67N Mn/SD	76Y Mn/SD	88M Mn/SD	91A Mn/SD	94B Mn/SD	Avg. Mn/SD
Percent Outstanding								
2. Crew Served Wpns	14/16	11/09	.	.	.	.	.	16/14
3. Tactical Mvmnts	17/18	08/10	.	.	.	.	.	13/14
4. Navigate	.	.	.	.	13/12	.	.	13/12
5. First Aid	.	.	.	.	.	12/14	.	12/14
7. Detect Targets	26/24	10/08	08/05	.	.	.	.	16/13
8. Repair Mech. Sys.	14/17	08/08	12/11	.	15/18	.	.	13/15
10. Use Tech Refs.	.	.	.	14/17	.	.	.	14/17
11. Pack and Load	15/17	09/11	.	16/20	14/11	.	11/07	18/17
13. Operate/Install	.	.	13/16	.	.	.	10/07	12/12
15. Operate Vehicles	18/20	10/11	.	.	19/19	.	.	18/20
16. Type	.	.	.	15/18	.	.	.	15/18
17. Record Keeping	.	.	08/08	20/23	11/07	11/13	.	13/13
18. Oral Comm.	17/19	09/11	.	.	.	13/14	.	15/16
19. Written Comm.	.	.	.	16/19	.	12/16	.	14/18
22. Medical Treatmnt.	.	.	.	.	.	12/14	.	12/14
23. Food Preparation	.	.	.	.	.	.	13/17	13/17
24. Leadership	.	.	.	.	.	13/17	.	13/17
Average	17/19	09/11	12/10	16/19	14/13	18/13	11/10	15/14
Sample Size	563	378	162	235	250	342	129	1807

TABLE 8.2
SOLDIER METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY TYPE OF JUDGE AND MOS

Level/	16S	19K	67N	76Y	88M	91A	94B	Avg.
Type of Judge	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD
Percent Unacceptable								
All Judges	11/16	09/09	12/13	18/21	15/16	14/15	13/13	13/15
FORSCOM Judges	12/16	09/10	12/13	23/23	16/16	13/14	12/10	14/15
TRADOC Judges	10/16	09/08	11/14	12/15	14/16	15/16	16/16	13/16
Officer Sessions	07/13	08/07	12/13	11/13	18/19	13/14	16/14	11/13
NCO Sessions	14/17	10/10	12/14	23/24	13/14	14/16	11/11	14/15
Percent Less Than Acceptable								
All Judges	29/22	26/21	28/20	41/27	34/22	37/24	34/21	33/23
FORSCOM Judges	30/22	25/21	29/20	47/27	37/22	38/26	33/22	35/23
TRADOC Judges	25/24	29/19	28/20	34/25	31/21	35/21	36/20	32/22
Officer Sessions	25/20	22/17	28/19	33/21	40/24	39/26	34/21	30/20
NCO Sessions	31/24	30/22	29/20	48/29	31/20	35/22	34/21	35/23
Percent Outstanding								
All Judges	16/19	09/10	11/12	14/19	14/15	11/15	10/09	12/15
FORSCOM Judges	17/19	11/12	11/14	14/20	13/13	14/18	09/08	13/15
TRADOC Judges	14/19	08/09	11/10	14/18	16/16	08/06	13/09	13/13
Officer Sessions	12/16	10/08	08/05	13/17	12/10	13/18	09/07	12/13
NCO Sessions	19/20	09/11	13/15	15/20	16/17	10/10	11/10	14/15

TABLE 8.3
SOLDIER METHOD:
SINGLE RATER RELIABILITY ESTIMATES
BY TYPE OF JUDGE AND ACCEPTABILITY LEVEL

	No. of	Acceptability Level			
Type of Judge	Obser-vations*	Unaccept-able	Less Than Acceptable	Out-standing	Avg.
All Judges	2052	.07	.09	.06	.07
FORSCOM Judges	1261	.11	.13	.06	.10
TRADOC Judges	791	.06	.08	.10	.08
NCO Sessions	1177	.09	.11	.10	.10
Officer Sessions	875	.13	.15	.04	.11

*Note: Each combination of judge and performance dimension is an observation.

Analyses of the Incident Method Data

Editing steps. The editing task for this instrument consisted of flagging missing data. No real editing was done; these cases were left as missing. There were 119 records out of 2208 (five percent) that contained missing data. Within these 119 records, a total of 294 responses were missing.

In addition to the above check, we also examined item-level data to identify particular incidents about which there was the most disagreement (highest standard deviations across judges). Appendix T in Volume II shows the means and standard deviations for each incident used in each scale.

Analysis by performance dimension. Table 8.04 shows the means and standard deviations of the judges' ratings of the percent of soldiers performing at each acceptability level for each combination of performance dimension and MOS. There are some interesting differences both within and between MOS in the degree of leniency and harshness. For example, for 16S, the percent of soldiers performing in the Unacceptable category are several points higher for dimensions 2 (Operate Crew-served Weapons), 3 (Tactical Movements), and 7 (Detect Targets) than for 11 (Pack and Load), and 15 (Operate Vehicles), which may be due to the appropriateness of the dimensions chosen for the MOS. Also, it appears that the MOS 16S and 19K have more stringent standards (higher means in the Unacceptable and Less than Acceptable categories) than the other MOS, except for 94B, in which only two dimensions are rated.

Analysis by type of judge. Table 8.5 shows the mean ratings (averaged across different dimensions) for each type of judge and MOS. In these analyses and in the analyses of the Task Method, two different samples were used. The first sample consisted of all judges. The second sample was limited to those judges who participated in Delphi sessions for the Incident or Task Method. We used this second, matched, sample in comparing pre- and post-Delphi results, so as to eliminate the effect

TABLE 8.4
INCIDENT METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY DIMENSION AND MOS (TOTAL PRE-DELPHI SAMPLES)

Level/ Performance Dimension	16S Mn/SD	19K Mn/SD	67N Mn/SD	76Y Mn/SD	88M Mn/SD	91A Mn/SD	94B Mn/SD	Avg. Mn/SD
Percent Unacceptable								
2. Crew Served Wpns	33/17	30/14	.	.	.	.	.	32/16
3. Tactical Mvmnts	31/20	.	.	.	.	.	.	31/20
4. Navigate	.	.	.	.	19/17	.	.	19/17
5. First Aid	.	.	.	.	.	27/21	.	27/21
7. Detect Targets	48/16	.	.	.	.	.	.	48/16
8. Repair Mech. Sys.	.	19/09	18/12	.	30/20	.	.	22/14
10. Use Tech Refs.	.	.	.	23/14	.	.	.	23/14
11. Pack and Load	27/20	.	.	.	.	.	30/22	29/21
13. Operate/Install	.	.	18/13	.	.	.	22/17	20/15
15. Operate Vehicles	17/18	25/21	.	.	12/07	.	.	18/15
16. Type	.	.	.	11/09	.	.	.	11/09
17. Record Keeping	.	.	15/13	20/17	18/13	.	.	18/14
18. Oral Comm.	16/18	20/17	.	.	.	15/14	.	17/16
19. Written Comm.	.	.	.	16/13	.	29/15	.	23/14
22. Medical Treatmnt.	.	.	.	.	.	17/13	.	17/13
Average	29/18	24/15	17/13	18/13	20/14	22/16	26/20	22/16
Sample Size	426	212	156	200	208	228	88	1518
Percent Less Than Acceptable								
2. Crew Served Wpns	51/18	46/18	.	.	.	.	.	49/18
3. Tactical Mvmnts	41/23	.	.	.	.	.	.	41/22
4. Navigate	.	.	.	.	35/25	.	.	35/25
5. First Aid	.	.	.	.	.	34/20	.	34/20
7. Detect Targets	53/16	.	.	.	.	.	.	53/16
8. Repair Mech. Sys.	.	23/12	26/20	.	31/19	.	.	27/17
10. Use Tech Refs.	.	.	.	27/18	.	.	.	27/18
11. Pack and Load	34/25	.	.	.	.	.	33/27	34/26
13. Operate/Install	.	.	21/15	.	.	.	30/18	26/17
15. Operate Vehicles	24/23	30/23	.	.	22/19	.	.	25/22
16. Type	.	.	.	17/15	.	.	.	17/15
17. Record Keeping	.	.	18/11	26/18	26/19	.	.	23/16
18. Oral Comm.	22/21	25/18	.	.	.	20/16	.	22/18
19. Written Comm.	.	.	.	27/16	.	37/13	.	32/15
22. Medical Treatmnt.	.	.	.	.	.	22/14	.	22/14
Average	38/21	31/18	22/15	24/17	29/21	28/16	32/23	29/19
Sample Size	426	212	156	200	208	228	88	1518

(Continued)

TABLE 8.4 (CONTINUED)
INCIDENT METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY DIMENSION AND MOS (TOTAL PRE-DELPHI SAMPLES)

Level/	16S	19K	67N	76Y	88M	91A	94B	Avg.
Performance Dimension	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD
Percent Outstanding								
2. Crew Served Wpns	14/14	15/14	.	.	.	.	.	15/14
3. Tactical Mvmnts	10/15	.	.	.	.	.	.	10/15
4. Navigate	.	.	.	.	11/11	.	.	11/11
5. First Aid	.	.	.	.	.	30/19	.	30/19
7. Detect Targets	19/20	.	.	.	.	.	.	19/20
8. Repair Mech. Sys.	.	13/13	20/15	.	21/19	.	.	18/16
10. Use Tech Refs.	.	.	.	21/20	.	.	.	21/20
11. Pack and Load	12/18	.	.	.	.	.	16/16	14/17
13. Operate/Install	.	.	17/14	.	.	.	23/20	20/17
15. Operate Vehicles	14/18	24/24	.	.	15/20	.	.	18/21
16. Type	.	.	.	17/17	.	.	.	17/17
17. Record Keeping	.	.	11/14	15/15	13/14	.	.	13/14
18. Oral Comm.	25/23	38/24	.	.	.	24/23	.	29/23
19. Written Comm.	.	.	.	15/16	.	18/23	.	17/20
22. Medical Treatmnt.	.	.	.	.	.	16/20	.	16/20
Average	16/18	23/19	16/14	17/17	15/16	22/21	20/18	18/18
Sample Size	426	212	156	200	208	228	88	1518

TABLE 8.5
INCIDENT METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY TYPE OF JUDGE, SAMPLE, AND MOS

Level/ Type of Judge/Sample	16S Mn/SD	19K Mn/SD	67N Mn/SD	76Y Mn/SD	88M Mn/SD	91A Mn/SD	94B Mn/SD	Avg. Mn/SD
Percent Unacceptable								
All Total Pre-Delphi	26/21	23/17	17/13	18/14	20/17	22/17	26/20	22/17
All Matched Pre-Delphi	26/21	22/18	17/14	17/14	20/17	22/17	22/15	21/17
All Matched Pst-Delphi	25/18	24/17	16/08	16/12	30/18	22/17	28/16	23/15
FORSCOM Total Pre-	24/20	19/14	15/12	18/14	19/17	21/16	24/20	20/16
FORSCOM Matched Pre-	25/21	17/12	14/07	16/09	21/19	20/14	20/12	19/13
FORSCOM Matched Pst-	23/18	22/18	16/06	15/08	36/18	18/13	21/11	22/13
TRADOC Total Pre-	32/23	29/18	19/13	17/15	21/16	23/17	29/19	24/17
TRADOC Matched Pre-	28/17	29/21	18/16	18/17	19/12	26/19	25/17	23/17
TRADOC Matched Pst-	29/18	27/15	16/09	17/14	20/12	27/20	36/16	25/15
Officer Total Pre-	23/16	22/14	18/13	16/10	17/11	20/14	24/17	20/14
Officer Matched Pre-	24/16	17/12	18/16	16/09	19/12	20/14	20/12	19/13
Officer Matched Pst-	24/16	22/18	16/09	15/08	20/12	18/13	21/11	19/13
NCO Total Pre-	29/23	25/18	16/12	19/17	22/19	23/19	27/22	23/19
NCO Matched Pre-	29/24	29/21	14/07	18/17	21/19	26/19	25/17	23/18
NCO Matched Pst-	25/20	27/15	16/06	17/14	36/18	27/20	36/16	26/16
Percent Less Than Acceptable								
All Total Pre-Delphi	35/25	31/20	22/16	24/17	28/21	28/17	32/23	29/20
All Matched Pre-Delphi	34/23	29/19	21/15	23/16	28/21	29/18	29/24	28/19
All Matched Pst-Delphi	32/21	31/19	19/08	22/15	36/19	26/16	31/18	28/17
FORSCOM Total Pre-	33/24	26/19	19/14	24/17	28/21	27/16	31/22	27/19
FORSCOM Matched Pre-	34/24	23/16	15/08	20/11	30/23	27/15	27/21	25/17
FORSCOM Matched Pst-	32/22	26/21	18/11	19/11	42/17	24/13	25/17	27/16
TRADOC Total Pre-	41/26	38/19	25/17	24/17	29/21	29/19	34/25	31/21
TRADOC Matched Pre-	35/19	37/20	23/17	25/19	25/17	32/20	31/27	30/20
TRADOC Matched Pst-	34/20	37/15	19/17	25/18	26/19	29/19	38/17	30/16
Officer Total Pre-	29/20	29/18	23/15	21/13	22/15	28/16	27/21	26/17
Officer Matched Pre-	30/20	23/16	23/17	20/11	25/17	27/15	27/21	25/17
Officer Matched Pst-	32/20	26/21	19/17	19/11	26/19	24/13	25/17	24/17
NCO Total Pre-	40/27	33/21	21/17	26/20	32/23	29/19	35/25	31/22
NCO Matched Pre-	38/25	37/20	15/08	25/19	30/23	32/20	31/27	30/20
NCO Matched Pst-	32/23	37/15	18/11	25/18	42/17	29/19	38/17	32/17

(Continued)

TABLE 8.5 (CONTINUED)
INCIDENT METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY TYPE OF JUDGE, SAMPLE, AND MOS

Level/ Type of Judge/Sample	16S Mn/SD	19K Mn/SD	67N Mn/SD	76Y Mn/SD	88M Mn/SD	91A Mn/SD	94B Mn/SD	Avg. Mn/SD
Percent Outstanding								
All Total Pre-Delphi	16/19	22/22	16/15	17/17	15/17	22/22	20/18	18/19
All Matched Pre-Delphi	16/19	22/20	14/12	20/19	16/18	23/23	19/17	19/18
All Matched Pst-Delphi	12/15	16/19	13/12	20/20	24/20	17/18	16/16	17/17
FORSCOM Total Pre-	16/19	15/14	13/12	16/16	15/17	21/20	17/13	19/16
FORSCOM Matched Pre-	16/20	15/12	16/12	17/16	17/18	19/17	18/16	17/16
FORSCOM Matched Pst-	12/16	13/12	16/12	17/16	32/21	15/14	10/10	16/14
TRADOC Total Pre-	15/19	32/26	19/16	18/19	15/17	23/24	24/24	21/21
TRADOC Matched Pre-	14/15	29/24	14/12	23/22	15/18	29/28	20/18	21/20
TRADOC Matched Pst-	11/14	20/25	12/12	24/23	11/10	21/22	22/20	17/18
Officer Total Pre-	12/14	23/23	12/12	15/14	13/15	18/17	24/22	17/17
Officer Matched Pre-	12/13	15/12	14/12	17/16	15/18	19/17	18/16	16/15
Officer Matched Pst-	10/12	13/12	12/12	17/16	11/10	15/14	10/10	13/12
NCO Total Pre-	18/21	22/21	20/16	18/19	16/18	25/25	16/13	19/19
NCO Matched Pre-	19/23	29/24	16/12	23/22	17/18	29/28	20/18	22/21
NCO Matched Pst-	14/18	20/25	16/12	24/23	32/21	21/22	22/20	21/20

of differences between the total sample and the subsample to whom the Delphi sessions were administered.

There are some differences between the ratings provided at FORSCOM sites versus those provided at TRADOC sites. By and large, the ratings obtained at FORSCOM sites appear to be slightly more lenient, with smaller means in the Unacceptable and Less than Acceptable categories, than those obtained at TRADOC sites. Likewise, the officers appear to be slightly more lenient than the NCOs. These results must be interpreted with caution, however, given the large standard deviations.

Table 8.6 shows estimates of single-rater reliability for each type of judge and each acceptability level. These results indicate that in the pre-Delphi condition, the NCOs' ratings were less reliable than the officers', but in the post-Delphi, the reliabilities were approximately equal. This suggests that the Delphi technique is of great benefit to the NCOs. If, in operation, NCOs are to provide standards using the Incident method, use of the Delphi technique should strongly be considered. No data are available for mixed groups (officers and NCOs), so the impact of the Delphi approach for such groups is unknown. The pre-Delphi ratings at obtained at FORSCOM locations were somewhat lower than those obtained at TRADOC locations, yet the reverse is true for the post-Delphi ratings. Neither of these differences is large.

Analyses of the Task-HS Method Data

Editing steps. We began by examining the ratings of the hypothetical soldier, flagging any cases where one soldier with a lower percent-GO score than some other soldier was judged to be in a higher acceptability category than the other soldier. We found 170 out of 3260 records, approximately five percent, with this type of error and resolved them by setting the record to missing.

TABLE 8.6
INCIDENT METHOD:
SINGLE RATER RELIABILITY ESTIMATES
BY TYPE OF JUDGE, SAMPLE, AND ACCEPTABILITY LEVEL

	No. of	Acceptability Level			
Type of Judge	Obser-vations*	Unaccept-able	Less Than Acceptable	Out-standing	Avg.
All Judges					
Total Pre-Delphi	1518	.18	.20	.11	.17
Matched Pre-Delphi	746	.19	.19	.11	.16
Matched Pst-Delphi	746	.31	.34	.18	.28
FORSCOM Judges					
Total Pre-Delphi	919	.25	.25	.10	.21
Matched Pre-Delphi	454	.27	.28	.11	.22
Matched Pst-Delphi	454	.49	.50	.37	.45
TRADOC Judges					
Total Pre-Delphi	599	.26	.27	.23	.25
Matched Pre-Delphi	292	.33	.31	.23	.29
Matched Pst-Delphi	292	.40	.42	.29	.37
NCO Sessions					
Total Pre-Delphi	859	.18	.18	.10	.16
Matched Pre-Delphi	364	.23	.22	.16	.20
Matched Pst-Delphi	364	.41	.35	.36	.37
Officer Sessions					
Total Pre-Delphi	659	.27	.31	.17	.27
Matched Pre-Delphi	382	.44	.42	.15	.34
Matched Pst-Delphi	382	.48	.57	.13	.37

*Note: Each combination of judge and performance dimension is an observation.

Analysis by performance dimension. Table 8.7 shows the means and standard deviations of the judges ratings of the percent of soldiers performing at each acceptability level for each combination of performance dimension and MOS. Using this standard setting technique, the 16S raters and the 19K raters appear to provide more stringent standards than do the other MOS. At the Less than Acceptable and Outstanding levels there are very few MOS differences.

Analysis by type of judge. Table 8.08 shows the mean ratings (averaged across different dimensions) for each type of judge and MOS. In the Unacceptable category, there are fairly large differences, approximately 20 points, between the ratings provided by TRADOC and FORSCOM judges in the 16S MOS. The difference is not nearly as great in the other MOS. This difference is noticeable at the Less than Acceptable level for 16S, but less extreme, approximately 10-15 points. There are very few interpretable differences in these data since the standard deviations are quite high.

Table 8.9 shows estimates of single-rater reliability for each type of judge and each acceptability level. One fairly large difference here is between FORSCOM and TRADOC ratings in the Less than Acceptable category in which the FORSCOM ratings are more reliable. Generally, the Officers' ratings are more reliable than the NCOs' ratings and both sets of ratings improve following the Delphi technique.

Analyses of the Task-DPG Method Data

Editing steps. For both the Task-DPG and the Task-APG data, we checked to see that the minimum percent-GO for the Marginal category was less than the minimum percent-GO for the Acceptable category and that the minimum for Acceptable was less than the minimum for Outstanding.

TABLE 8.7
TASK-HS METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY DIMENSION AND MOS (TOTAL PRE-DELPHI SAMPLES)

Level/	16S	19K	67N	76Y	88M	91A	Avg.
Performance Dimension	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD
Percent Unacceptable							
2. Crew Served Wpns	38/24	54/28	.	.	.	.	46/26
4. Navigate	.	.	.	.	23/11	.	23/11
5. First Aid	.	.	.	.	.	33/30	33/30
8. Repair Mech. Sys.	.	21/30	14/11	.	12/15	.	16/15
15. Operate Vehicles	34/26	43/31	43/19	.	26/20	.	37/24
16. Type	.	.	.	31/23	.	.	31/23
17. Record Keeping	.	.	37/21	41/19	27/20	43/21	37/20
18. Oral Comm.	42/24	41/26	.	.	.	36/21	40/24
19. Written Comm.	.	.	.	25/11	.	18/18	22/15
22. Medical Treatmnt.	.	.	.	.	.	46/35	46/35
Average	38/25	40/26	31/17	32/18	22/17	35/25	33/21
Sample Size	209	148	67	75	102	146	747
Percent Less Than Acceptable							
2. Crew Served Wpns	58/28	75/24	.	.	.	.	67/26
4. Navigate	.	.	.	.	52/13	.	52/13
5. First Aid	.	.	.	.	.	58/32	58/32
8. Repair Mech. Sys.	.	43/36	38/18	.	32/26	.	44/28
15. Operate Vehicles	57/29	65/31	58/22	.	59/24	.	60/28
16. Type	.	.	.	64/19	.	.	64/19
17. Record Keeping	.	.	64/19	65/16	57/21	72/11	65/17
18. Oral Comm.	58/29	57/31	.	.	.	59/24	58/28
19. Written Comm.	.	.	.	52/16	.	44/20	48/18
22. Medical Treatmnt.	.	.	.	.	.	72/29	72/19
Average	58/27	60/31	61/21	60/17	50/21	61/21	59/23
Sample Size	209	148	67	75	102	146	747

(Continued)

TABLE 8.7 (CONTINUED)
TASK-HS METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY DIMENSION AND MOS (TOTAL PRE-DELPHI SAMPLES)

Level/	16S	19K	67N	76Y	88M	91A	Avg.
Performance Dimension	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD
Percent Outstanding							
2. Crew Served Wpns	06/06	08/16	.	.	.	.	07/11
4. Navigate	.	.	.	.	16/12	.	16/12
5. First Aid	.	.	.	.	.	07/09	07/09
8. Repair Mech. Sys.	.	14/25	21/11	.	16/15	.	17/17
15. Operate Vehicles	06/07	09/16	10/06	.	09/10	.	09/10
16. Type	.	.	.	04/09	.	.	04/09
17. Record Keeping	.	.	08/06	07/06	09/06	05/06	07/06
18. Oral Comm.	06/11	12/12	.	.	.	08/11	09/11
19. Written Comm.	.	.	.	14/08	.	14/10	14/09
22. Medical Treatmnt.	.	.	.	.	.	02/04	02/04
Average	06/08	11/17	13/08	08/08	13/11	07/08	09/11
Sample Size	209	148	67	75	102	146	747

TABLE 8.8
TASK-HS METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY TYPE OF JUDGE, SAMPLE, AND MOS

Level/ Type of Judge/Sample	16S Mn/SD	19K Mn/SD	67N Mn/SD	76Y Mn/SD	88M Mn/SD	91A Mn/SD	Avg. Mn/SD
Percent Unacceptable							
All Total Pre-Delphi	39/25	40/31	32/21	34/19	22/18	36/29	34/24
All Matched Pre-Delphi	35/26	53/23	36/22	42/20	24/17	28/23	36/22
All Matched Pst-Delphi	44/28	67/27	40/19	47/20	34/17	27/22	43/22
FORSCOM Total Pre-	33/21	47/29	33/22	34/19	18/17	33/24	33/22
FORSCOM Matched Pre-	22/17	53/23	35/23	42/20		29/25	36/22
FORSCOM Matched Pst-	33/15	67/27	45/17	47/20		26/22	44/22
TRADOC Total Pre-	53/28	34/31	31/21	32/17	26/18	41/36	36/25
TRADOC Matched Pre-	52/27		37/23		24/17	26/19	35/22
TRADOC Matched Pst-	58/35		37/20		34/17	31/22	40/24
Officer Total Pre-	43/25	45/31	32/21	34/19	28/17	34/23	36/23
Officer Matched Pre-	37/17		35/23			26/19	33/20
Officer Matched Pst-	38/15		45/17			31/22	38/18
NCO Total Pre-	36/25	36/31	31/21	33/18	18/17	38/34	32/24
NCO Matched Pre-	34/28	53/23	37/23	42/20	24/17	29/25	37/23
NCO Matched Pst-	45/30	67/27	37/20	47/20	34/17	26/22	43/23
Percent Less Than Acceptable							
All Total Pre-Delphi	58/29	60/33	56/22	61/17	49/24	62/28	58/26
All Matched Pre-Delphi	54/30	69/26	60/21	65/19	55/19	56/24	60/23
All Matched Pst-Delphi	63/26	79/26	64/16	68/21	62/13	59/25	56/18
FORSCOM Total Pre-	55/28	68/31	61/20	60/17	43/25	63/24	58/24
FORSCOM Matched Pre-	48/29	69/26	62/18	65/19		57/24	60/23
FORSCOM Matched Pst-	59/16	79/26	66/14	68/21		57/27	66/21
TRADOC Total Pre-	67/30	55/33	53/23	63/18	57/23	59/34	59/27
TRADOC Matched Pre-	64/30		59/23		55/19	54/22	58/24
TRADOC Matched Pst-	69/34		62/17		62/13	64/20	64/21
Officer Total Pre-	64/27	66/33	55/23	63/16	56/22	65/23	62/24
Officer Matched Pre-	63/20		62/18			54/22	60/20
Officer Matched Pst-	66/16		66/14			64/20	65/17
NCO Total Pre-	54/29	56/32	56/21	58/18	45/25	59/31	55/26
NCO Matched Pre-	53/32	69/26	59/23	65/19	55/19	57/24	60/24
NCO Matched Pst-	63/27	79/26	62/17	68/21	62/13	57/27	65/22

(Continued)

TABLE 8.8 (CONTINUED)
TASK-HS METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY TYPE OF JUDGE, SAMPLE, AND MOS

Level/ Type of Judge/Sample	16S Mn/SD	19K Mn/SD	67N Mn/SD	76Y Mn/SD	88M Mn/SD	91A Mn/SD	Avg. Mn/SD
Percent Outstanding							
All Total Pre-Delphi	06/08	11/18	11/10	09/08	12/12	06/09	09/11
All Matched Pre-Delphi	06/10	07/09	08/07	07/06	12/11	09/10	08/09
All Matched Pst-Delphi	05/07	01/04	09/06	05/05	10/07	07/09	06/06
FORSCOM Total Pre-	07/09	06/09	10/10	08/08	14/13	07/09	09/10
FORSCOM Matched Pre-	08/12	07/09	09/07	07/06		10/11	08/09
FORSCOM Matched Pst-	08/08	01/04	05/05	05/05		07/08	05/05
TRADOC Total Pre-	03/05	14/21	12/09	12/07	11/11	05/08	10/10
TRADOC Matched Pre-	03/05		08/07		12/11	08/07	08/08
TRADOC Matched Pst-	02/04		12/05		10/07	08/09	08/06
Officer Total Pre-	06/08	05/08	13/10	08/08	11/10	06/07	08/09
Officer Matched Pre-	11/15		09/07			08/07	09/10
Officer Matched Pst-	07/10		05/05			08/09	07/08
NCO Total Pre-	06/09	15/21	10/09	10/08	14/13	07/10	10/12
NCO Matched Pre-	05/08	07/09	08/07	07/06	12/11	10/11	08/09
NCO Matched Pst-	05/07	01/04	12/05	05/05	10/07	07/08	07/06

We also checked for missing values and found 22 cases where the minimum for Outstanding was missing. In most of these cases, the minimum for Acceptable was 90 percent-GO or better so it was logical to assume that the rater felt that even if a soldier scored 100 percent, he/she should not be considered Outstanding. In such cases, we inserted a value of 100 for the Outstanding minimum.

In cases where there were extreme values, such as Marginal = 5%, Acceptable = 6% and Outstanding = 7% or if there were missing data in the marginal and/or acceptable categories, the remaining values for the record were set to missing. If there was only one task that passed the above error screens, then the record was deleted. There was a total of 132 out of 978 records, 13.5 percent, that had at least one of the above errors.

We also examined the ratings for each individual task, even though all of the subsequent analyses used the overall percent-GO scores for the dimension as a whole. Volume II, Appendix U shows the means and standard deviations of the judgments for the individual tasks. These data will be used to suggest revisions to the set of tasks used to illustrate each dimension.

Analysis by performance dimension. Table 8.10 shows the means and standard deviations of the judges ratings of the percent of soldiers performing at each acceptability level for each combination of performance dimension and MOS. As with the Task-HS method, 16S and 19K appear to be the most harsh, with higher ratings in the Unacceptable and Less than Acceptable categories. Within MOS, there do not appear to be any differences among the dimensions that are large enough to be significant given the large standard deviations.

Analysis by type of judge. Table 8.11 shows the mean ratings (averaged across different dimensions) for each type of judge and MOS. The ratings at both TRADOC and FORSCOM posts became more harsh in all three performance categories. Ratings from 16S raters at TRADOC posts are more harsh in their post-Delphi ratings than 16S

TABLE 8.9
TASK-HS METHOD:
SINGLE RATER RELIABILITY ESTIMATES
BY TYPE OF JUDGE, SAMPLE, AND ACCEPTABILITY LEVEL

		Acceptability Level			
Type of Judge	No. of Observations*	Unacceptable	Less Than Acceptable	Outstanding	Avg.
All Judges					
Total Pre-Delphi	747	.15	.12	.13	.13
Matched Pre-Delphi	231	.14	.08	.12	.11
Matched Pst-Delphi	231	.22	.09	.28	.20
FORSCOM Judges					
Total Pre-Delphi	445	.22	.20	.16	.19
Matched Pre-Delphi	131	.24	.16	.08	.16
Matched Pst-Delphi	131	.38	.15	.28	.27
TRADOC Judges					
Total Pre-Delphi	302	.24	.12	.18	.18
Matched Pre-Delphi	100	.27	.06	.35	.23
Matched Pst-Delphi	100	.24	.08	.48	.27
NCO Sessions					
Total Pre-Delphi	411	.15	.11	.19	.15
Matched Pre-Delphi	188	.13	.09	.16	.13
Matched Pst-Delphi	188	.20	.10	.33	.21
Officer Sessions					
Total Pre-Delphi	336	.26	.23	.28	.26
Matched Pre-Delphi	43	.24	.16	.14	.18
Matched Pst-Delphi	43	.44	.22	.27	.31

*Note: Each combination of judge and performance dimension is an observation.

TABLE 8.10
TASK-DPG METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY DIMENSION AND MOS (TOTAL PRE-DELPHI SAMPLES)

Level/ Performance Dimension	16S Mn/SD	19K Mn/SD	67N Mn/SD	76Y Mn/SD	88M Mn/SD	91A Mn/SD	Avg. Mn/SD
Percent Unacceptable							
2. Crew Served Wpns	43/17	46/23	.	.	.	.	45/20
4. Navigate	.	.	.	.	36/13	.	36/13
5. First Aid	.	.	.	.	.	34/23	34/23
8. Repair Mech. Sys.	.	23/19	19/08	.	21/10	.	21/12
13. Operate/Install	.	.	25/09	.	.	.	25/09
15. Operate Vehicles	31/17	39/29	.	.	32/19	.	34/22
16. Type	.	.	.	34/21	.	.	34/21
17. Record Keeping	.	.	45/15	48/14	41/16	52/15	47/15
18. Oral Comm.	59/17	59/28	.	.	.	45/16	54/20
19. Written Comm.	.	.	.	23/11	.	21/19	22/15
22. Medical Treatmnt.	.	.	.	.	.	48/32	48/32
Average	44/17	42/25	30/11	35/15	33/15	40/21	37/17
Sample Size	180	121	62	71	95	134	663
Percent Less Than Acceptable							
2. Crew Served Wpns	66/15	69/23	.	.	.	.	68/19
4. Navigate	.	.	.	.	58/12	.	58/12
5. First Aid	.	.	.	.	.	60/19	60/19
8. Repair Mech. Sys.	.	49/27	41/19	.	40/21	.	43/22
13. Operate/Install	.	.	51/19	.	.	.	51/19
15. Operate Vehicles	61/18	64/27	.	.	61/18	.	62/21
16. Type	.	.	.	61/16	.	.	61/16
17. Record Keeping	.	.	68/14	70/12	56/19	74/13	69/15
18. Oral Comm.	82/13	77/23	.	.	.	68/13	76/16
19. Written Comm.	.	.	.	49/22	.	50/21	50/22
22. Medical Treatmnt.	.	.	.	.	.	76/23	76/23
Average	70/15	65/25	53/17	60/17	56/18	66/18	62/18
Sample Size	180	121	62	71	95	134	663

(Continued)

TABLE 8.10 (CONTINUED)
TASK-DPG METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY DIMENSION AND MOS (TOTAL PRE-DELPHI SAMPLES)

Level/	16S	19K	67N	76Y	88M	91A	Avg.
Performance Dimension	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD
Percent Outstanding							
2. Crew Served Wpns	08/09	11/14	.	.	.	.	10/12
4. Navigate	.	.	.	.	12/09	.	12/09
5. First Aid	.	.	.	.	.	09/09	09/09
8. Repair Mech. Sys.	.	19/24	20/07	.	21/19	.	20/17
13. Operate/Install	.	.	17/21	.	.	.	17/12
15. Operate Vehicles	06/05	12/17	.	.	10/13	.	09/12
16. Type	.	.	.	06/07	.	.	06/07
17. Record Keeping	.	.	05/03	05/04	08/06	06/06	06/05
18. Oral Comm.	03/03	09/13	.	.	.	05/05	06/07
19. Written Comm.	.	.	.	15/19	.	14/10	15/15
22. Medical Treatmnt.	.	.	.	.	.	04/08	04/08
Average	06/06	13/17	14/10	09/10	13/12	08/07	11/10
Sample Size	180	121	62	71	95	134	663

TABLE 8.11
TASK-DPG METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY TYPE OF JUDGE, SAMPLE, AND MOS

Level/ Type of Judge/Sample	16S Mn/SD	19K Mn/SD	67N Mn/SD	76Y Mn/SD	88M Mn/SD	91A Mn/SD	Avg. Mn/SD
Percent Unacceptable							
All Total Pre-Delphi	47/21	42/28	35/18	38/18	32/16	40/26	39/21
All Matched Pre-Delphi	45/23	70/24	38/17	51/12	40/16	32/23	46/19
All Matched Pst-Delphi	51/26	80/17	40/20	55/12	40/11	32/20	50/18
FORSCOM Total Pre-	42/19	52/26	39/17	38/19	27/16	38/23	39/20
FORSCOM Matched Pre-	32/16	70/24	43/14	51/12		35/26	46/18
FORSCOM Matched Pst-	38/19	80/17	49/15	55/12		32/21	51/17
TRADOC Total Pre-	57/20	33/37	32/17	38/18	37/16	45/30	40/21
TRADOC Matched Pre-	62/18		36/18		40/16	27/14	41/17
TRADOC Matched Pst-	69/24		35/21		40/11	32/18	44/19
Officer Total Pre-	45/21	46/23	36/18	40/18	35/15	37/20	40/19
Officer Matched Pre-	37/18		68/11			27/14	36/15
Officer Matched Pst-	38/18		49/15			32/18	40/17
NCO Total Pre-	47/20	40/30	33/18	35/18	29/17	44/31	38/22
NCO Matched Pre-	47/23	70/24	36/18	51/12	40/16	35/26	47/20
NCO Matched Pst-	54/27	80/17	35/21	55/12	40/11	32/21	49/18
Percent Less Than Acceptable							
All Total Pre-Delphi	71/18	65/27	58/20	62/19	55/21	67/21	63/21
All Matched Pre-Delphi	70/19	86/19	62/17	72/08	65/13	60/22	69/16
All Matched Pst-Delphi	76/19	93/08	62/17	74/13	65/11	62/20	72/15
FORSCOM Total Pre-	69/17	75/21	65/18	61/19	50/20	66/21	64/19
FORSCOM Matched Pre-	62/19	86/19	68/11	72/08		61/24	70/16
FORSCOM Matched Pst-	66/18	93/08	69/12	74/13		62/21	73/14
TRADOC Total Pre-	77/17	56/28	52/20	65/17	61/20	68/22	63/21
TRADOC Matched Pre-	81/13		59/20		65/13	58/17	66/16
TRADOC Matched Pst-	89/12		58/19		65/11	62/19	69/15
Officer Total Pre-	70/19	72/21	58/21	64/16	56/21	68/18	65/19
Officer Matched Pre-	63/20		68/11			58/17	63/16
Officer Matched Pst-	70/16		69/12			62/19	67/16
NCO Total Pre-	72/17	60/29	58/20	58/22	54/20	66/25	61/22
NCO Matched Pre-	72/18	86/19	59/20	72/08	65/13	61/24	69/17
NCO Matched Pst-	77/20	93/08	58/19	74/13	65/11	62/21	72/15

(Continued)

TABLE 8.11 (CONTINUED) TASK-DPG METHOD: MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY TYPE OF JUDGE, SAMPLE, AND MOS

Level/	16S	19K	67N	76Y	88M	91A	Avg.
Type of Judge/Sample	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD
Percent Outstanding							
All Total Pre-Delphi	06/06	12/18	11/11	08/12	13/14	07/09	10/12
All Matched Pre-Delphi	05/05	06/14	09/14	03/02	10/08	08/07	07/08
All Matched Pst-Delphi	06/08	02/04	07/05	03/02	11/08	07/06	06/06
FORSCOM Total Pre-	06/07	07/10	09/08	08/13	13/12	08/10	09/10
FORSCOM Matched Pre-	07/06	06/14	06/03	03/02		08/08	06/07
FORSCOM Matched Pst-	08/10	02/04	05/03	03/02		07/06	05/05
TRADOC Total Pre-	03/05	18/22	13/14	09/06	13/16	06/07	10/12
TRADOC Matched Pre-	03/03		11/17		10/08	07/05	08/08
TRADOC Matched Pst-	02/03		08/07		11/08	06/06	07/06
Officer Total Pre-	06/06	06/07	10/08	07/06	14/17	06/06	08/08
Officer Matched Pre-	06/06		06/03			07/05	06/05
Officer Matched Pst-	03/03		05/03			06/06	05/04
NCO Total Pre-	05/07	17/21	12/14	10/17	13/12	08/11	11/14
NCO Matched Pre-	05/05	06/14	11/17	03/02	10/08	08/08	07/09
NCO Matched Pst-	06/09	02/04	08/07	03/02	11/08	07/06	06/06

raters at FORSCOM posts, whereas 67N post-Delphi ratings are more harsh when obtained at FORSCOM locations than at TRADOC locations. 91A post-Delphi ratings are nearly the same, regardless of locations. These MOS differences, although somewhat small given the standard deviations, hold for all three proficiency levels.

Table 8.12 shows estimates of single-rater reliability for each type of judge and each acceptability level. For ratings obtained in both FORSCOM and TRADOC locations, the reliabilities of the post-Delphi ratings are higher than that of the pre-Delphi ratings. The reliabilities of the NCO ratings increased following the Delphi technique at all three proficiency levels, whereas the reliabilities of the officers' ratings decreased slightly when compared with the matched pre-Delphi sample, but increased slightly when compared with the total sample.

Analyses of the Task-APG Method Data

Analysis by performance dimension. Table 8.13 shows the means and standard deviations of the judges' ratings of the percent of soldiers performing at each acceptability level for each combination of performance dimension and MOS. Since there was only Project A data available on ten dimensions, only a limited amount of data could be converted into the common metric for this instrument. Hence, the sample sizes are quite small. Again, given the relatively high standard deviations, meaningful interpretation of these data is difficult. There do not appear to be significant differences across MOS in terms of leniency/harshness.

Analysis by type of judge. Table 8.14 shows the mean ratings (averaged across different dimensions) for each type of judge and MOS. As with the other instruments, the post-Delphi results show more harshness than the pre-Delphi results. For 67N, the FORSCOM ratings are more harsh than the ratings obtained at TRADOC locations at all three levels of proficiency. For 88M and 91A, the results are similar.

TABLE 8.12
TASK-DPG METHOD:
SINGLE RATER RELIABILITY ESTIMATES
BY TYPE OF JUDGE, SAMPLE, AND ACCEPTABILITY LEVEL

	No. of	Acceptability Level			
Type of Judge	Obser-vations*	Unaccept-able	Less Than Acceptable	Out-standing	Avg.
All Judges					
Total Pre-Delphi	663	.29	.28	.18	.25
Matched Pre-Delphi	187	.30	.34	.21	.30
Matched Pst-Delphi	187	.47	.40	.26	.38
FORSCOM Judges					
Total Pre-Delphi	403	.38	.34	.21	.31
Matched Pre-Delphi	101	.48	.34	.23	.35
Matched Pst-Delphi	101	.64	.51	.31	.49
TRADOC Judges					
Total Pre-Delphi	260	.40	.35	.24	.33
Matched Pre-Delphi	86	.60	.45	.26	.44
Matched Pst-Delphi	86	.56	.56	.41	.51
NCO Sessions					
Total Pre-Delphi	359	.28	.30	.25	.28
Matched Pre-Delphi	146	.35	.27	.22	.28
Matched Pst-Delphi	146	.46	.40	.26	.37
Officer Sessions					
Total Pre-Delphi	304	.42	.39	.29	.37
Matched Pre-Delphi	41	.64	.51	.50	.55
Matched Pst-Delphi	41	.62	.42	.38	.47

*Note: Each combination of judge and performance dimension is an observation.

TABLE 8.13
TASK-APG METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY DIMENSION AND MOS (TOTAL PRE-DELPHI SAMPLES)

Level/ Performance Dimension	16S Mn/SD	19K Mn/SD	67N Mn/SD	76Y Mn/SD	88M Mn/SD	91A Mn/SD	Avg. Mn/SD
Percent Unacceptable							
2. Crew Served Wpns	43/11	55/16	.	.	.	.	49/14
4. Navigate	.	.	.	.	47/11	.	47/11
5. First Aid	.	.	.	.	.	28/23	28/23
8. Repair Mech. Sys.	.	26/21	24/20	.	25/14	.	25/18
15. Operate Vehicles	34/22	55/27	.	.	47/18	.	45/22
16. Type	.	.	.	55/24	.	.	55/24
17. Record Keeping	.	.	54/06	.	55/18	50/14	53/13
18. Oral Comm.	.	76/20	.	.	.	65/12	71/16
19. Written Comm.	.	.	.	27/17	.	25/15	26/16
22. Medical Treatmnt.	.	.	.	.	.	37/28	37/28
Average	39/17	53/21	39/13	41/21	44/15	41/18	43/18
Sample Size	41	56	31	44	74	109	355
Percent Less Than Acceptable							
2. Crew Served Wpns	66/10	76/12	.	.	.	.	71/11
4. Navigate	.	.	.	.	64/10	.	64/10
5. First Aid	.	.	.	.	.	54/23	54/23
8. Repair Mech. Sys.	.	51/26	43/24	.	44/20	.	46/23
15. Operate Vehicles	65/19	78/21	.	.	74/14	.	72/18
16. Type	.	.	.	74/24	.	.	74/24
17. Record Keeping	.	.	77/11	.	76/22	73/13	75/15
18. Oral Comm.	.	90/11	.	.	.	85/08	88/10
19. Written Comm.	.	.	.	47/22	.	46/19	47/21
22. Medical Treatmnt.	.	.	.	.	.	78/14	78/14
Average	66/15	74/18	60/18	61/23	65/17	67/15	66/18
Sample Size	41	56	31	44	74	109	355

(Continued)

TABLE 8.13 (CONTINUED)
TASK-APG METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY DIMENSION AND MOS (TOTAL PRE-DELPHI SAMPLES)

Level/ Performance Dimension	16S Mn/SD	19K Mn/SD	67N Mn/SD	76Y Mn/SD	88M Mn/SD	91A Mn/SD	Avg. Mn/SD
Percent Outstanding							
2. Crew Served Wpns	10/07	08/05	.	.	.	.	09/06
4. Navigate	.	.	.	.	11/06	.	11/06
5. First Aid	.	.	.	.	.	14/12	14/12
8. Repair Mech. Sys.	.	14/16	22/19	.	16/07	.	17/14
15. Operate Vehicles	05/10	04/05	.	.	05/05	.	05/07
16. Type	.	.	.	07/21	.	.	07/21
17. Record Keeping	.	.	07/08	.	04/05	07/08	06/07
18. Oral Comm.	.	02/02	.	.	.	03/03	03/03
19. Written Comm.	.	.	.	20/16	.	16/16	18/16
22. Medical Treatmnt.	.	.	.	.	.	02/01	02/01
Average	08/09	07/07	15/14	14/19	09/06	08/08	10/11
Sample Size	41	56	31	44	74	109	355

TABLE 8.14
TASK-APG METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY TYPE OF JUDGE, SAMPLE, AND MOS

Level/ Type of Judge/Sample	16S Mn/SD	19K Mn/SD	67N Mn/SD	76Y Mn/SD	88M Mn/SD	91A Mn/SD	Avg. Mn/SD
Percent Unacceptable							
All Total Pre-Delphi	37/19	54/27	29/21	39/24	46/17	41/24	35/22
All Matched Pre-Delphi	45/09	41/24	18/11	50/23	40/19	36/24	38/18
All Matched Pst-Delphi	60/14	67/32	19/12	48/20	39/16	53/29	48/21
FORSCOM Total Pre-	37/19	54/27	37/23	38/24	47/19	44/25	43/23
FORSCOM Matched Pre-	45/09	41/24	22/13	50/23	40/23	41/23	40/19
FORSCOM Matched Pst-	60/14	67/32	25/15	48/20	43/16	39/21	47/20
TRADOC Total Pre-			15/06	42/26	43/12	38/23	35/17
TRADOC Matched Pre-			15/06		39/11	28/25	27/14
TRADOC Matched Pst-			13/04		33/15	75/26	40/15
Officer Total Pre-	43/11	66/26	31/23	37/27	43/20	39/27	43/22
Officer Matched Pre-	45/09		22/13		40/23	28/25	34/18
Officer Matched Pst-	60/14		25/15		43/16	75/26	51/18
NCO Total Pre-	34/22	41/23	27/20	41/23	48/15	43/21	39/21
NCO Matched Pre-	41/24		15/06	50/23	39/11	41/23	37/17
NCO Matched Pst-	67/32		13/04	48/20	33/15	39/21	40/18
Percent Less Than Acceptable							
All Total Pre-Delphi	65/16	74/23	49/25	59/26	67/19	65/23	63/22
All Matched Pre-Delphi	66/10	64/24	37/18	70/26	64/21	64/22	61/20
All Matched Pst-Delphi	80/09	85/22	37/16	73/16	67/15	73/23	69/17
FORSCOM Total Pre-	65/16	74/23	59/24	57/28	67/21	66/24	65/23
FORSCOM Matched Pre-	66/10	64/24	44/21	70/26	61/25	66/20	62/21
FORSCOM Matched Pst-	80/09	85/22	46/17	73/16	68/17	62/22	69/17
TRADOC Total Pre-			30/12	63/23	67/12	63/22	56/17
TRADOC Matched Pre-			30/12		69/13	62/26	54/17
TRADOC Matched Pst-			28/07		67/13	89/16	61/12
Officer Total Pre-	66/10	84/18	53/25	54/28	62/21	65/26	64/21
Officer Matched Pre-	66/10		44/21		61/25	62/26	58/21
Officer Matched Pst-	80/09		46/17		68/17	89/16	71/15
NCO Total Pre-	65/19	64/23	44/25	63/25	70/17	65/20	62/22
NCO Matched Pre-	64/24		30/12	70/26	69/13	66/20	60/19
NCO Matched Pst-	85/22		28/07	73/16	67/13	62/22	63/16

(Continued)

TABLE 8.14 (CONTINUED)
TASK-APG METHOD:
MEAN/STANDARD DEVIATION OF PERCENT OF SOLDIERS AT EACH LEVEL BY TYPE OF JUDGE, SAMPLE, AND MOS

Level/	16S	19K	67N	76Y	88M	91A	Avg.
Type of Judge/Sample	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD	Mn/SD
Percent Outstanding							
All Total Pre-Delphi	07/09	07/09	19/19	14/20	08/07	10/12	11/13
All Matched Pre-Delphi	10/07	11/12	25/19	08/25	08/08	10/11	12/14
All Matched Pst-Delphi	04/03	02/06	22/13	02/02	07/06	07/09	07/07
FORSCOM Total Pre-	07/09	07/09	12/10	13/21	08/07	10/14	10/12
FORSCOM Matched Pre-	10/07	11/12	17/09	08/25	10/08	11/10	11/12
FORSCOM Matched Pst-	04/03	02/06	17/09	02/02	07/07	10/11	07/06
TRADOC Total Pre-			33/23	15/18	09/07	09/09	17/14
TRADOC Matched Pre-			33/23		05/05	10/12	16/13
TRADOC Matched Pst-			28/15		06/05	02/04	12/08
Officer Total Pre-	10/07	03/04	13/10	15/14	11/08	10/15	10/10
Officer Matched Pre-	10/07		17/09		10/08	10/12	12/09
Officer Matched Pst-	04/03		17/09		07/07	02/04	09/07
NCO Total Pre-	06/10	11/12	25/23	13/23	06/06	10/08	12/14
NCO Matched Pre-	11/12		33/23	08/25	05/05	11/10	14/15
NCO Matched Pst-	02/06		28/15	02/02	06/05	10/11	10/08

Table 8.15 shows estimates of single-rater reliability for each type of judge and each acceptability level. Also, the reliability of post-Delphi results was higher than that of the pre-Delphi results.

Comparison of Task-DPG and Task-APG results. The Task-DPG and Task-APG methods are identical in format. The only difference is that for the Task-DPG method, a great deal of information is provided about the particular steps (items) that are considered in computing the percent-GO scores. It is reasonable to ask whether this additional information led to different standards or different levels of agreement among judges. In other words, Did the extra information help judges to reach a common understanding or just confuse them?

Table 8.16 shows the means and standard deviations of the percent-GO scores that resulted from each method, rater group, and acceptability level. As can be seen from this table, the APG method usually led to slightly harsher ratings, but also very slightly smaller standard deviations than the DPG method. The differences were minimal at most.

Comparisons Across Methods

Table 8.17 presents summary statistics (means, standard deviations and single-rater reliabilities) for each of the standard setting methods, averaged across dimension and MOS. Where Delphi sessions were used, statistics for the matched samples are shown along with the statistics for the entire sample.

The first general conclusion to be drawn from this table is that all of the variations of the task instrument lead to very strict standards. The percent judged to be unacceptable ranged from 34 to 48 with the task-based methods in comparison to 16 and 22 for the Soldier and Incident Methods respectively. Similarly, the percent less than fully acceptable ranges from 56 to 72 for the Task-Based Methods compared to 31 and 29 for the Soldier and Incident Methods. Differences at the high end of the scale,

TABLE 8.15
TASK-APG METHOD:
SINGLE RATER RELIABILITY ESTIMATES
BY TYPE OF JUDGE, SAMPLE, AND ACCEPTABILITY LEVEL

	No. of	Acceptability Level			
	Obser-	Unaccept-	Less Than	Out-	
Type of Judge	vations*	able	Acceptable	standing	Avg.
All Judges					
Total Pre-Delphi	355	.42	.40	.23	.35
Matched Pre-Delphi	148	.32	.38	.22	.31
Matched Pst-Delphi	148	.46	.49	.46	.47
FORSCOM Judges					
Total Pre-Delphi	258	.39	.36	.19	.31
Matched Pre-Delphi	109	.32	.30	.10	.24
Matched Pst-Delphi	109	.47	.45	.45	.46
TRADOC Judges					
Total Pre-Delphi	97	.60	.63	.41	.55
Matched Pre-Delphi	39	.31	.64	.46	.47
Matched Pst-Delphi	39	.72	.82	.63	.72
NCO Sessions					
Total Pre-Delphi	194	.39	.37	.26	.34
Matched Pre-Delphi	87	.34	.37	.24	.32
Matched Pst-Delphi	87	.53	.56	.53	.54
Officer Sessions					
Total Pre-Delphi	161	.49	.50	.31	.43
Matched Pre-Delphi	61	.37	.47	.33	.39
Matched Pst-Delphi	61	.60	.64	.59	.61

*Note: Each combination of judge and performance dimension is an observation.

TABLE 8.16
COMPARISON OF TASK-BASED DPG AND APG PERCENT-GO BY TYPE OF JUDGE, SAMPLE, AND ACCEPTABILITY LEVEL

	Marginal		Acceptable		Outstanding	
	DPG	APG	DPG	APG	DPG	APG
ALL						
Total Pre-	66/12	69/10	78/09	80/08	92/06	93/06
Matched Pre-	65/11	69/09	78/08	80/08	92/06	92/06
Matched Pst-	67/12	72/11	80/09	83/08	93/05	94/05
FORSCOM						
Total Pre-	66/10	69/10	78/08	81/08	92/06	93/06
Matched Pre-	64/11	69/10	77/09	81/08	92/06	93/07
Matched Pst-	66/11	70/11	79/08	82/08	92/05	93/04
TRADOC						
Total Pre-	66/14	69/11	78/11	80/09	92/07	92/07
Matched Pre-	66/11	70/09	79/08	80/07	92/05	92/06
Matched Pst-	68/12	76/11	81/10	85/09	93/05	94/05
NCOs						
Total Pre-	65/13	68/11	78/10	79/09	91/07	92/07
Matched Pre-	66/11	69/10	78/09	80/08	92/06	92/07
Matched Pst-	68/12	70/12	80/10	82/09	93/05	93/05
OFFICERS						
Total Pre-	67/10	71/10	79/08	81/08	92/05	93/05
Matched Pre-	60/08	70/09	75/05	81/07	92/04	93/04
Matched Pst-	62/08	76/09	77/06	85/07	93/04	95/03

TABLE 8.17
SUMMARY OF RATING RESULTS BY JUDGMENT AND METHOD FOR TOTAL SAMPLE AND MATCHED PRE AND POST DELPHI SAMPLES

	Means			Standard Devs.			Reliabilities		
Level/	All	Match Smp		All	Match Smp		All	Match Smp	
Method	Pre	Pre	Post	Pre	Pre	Post	Pre	Pre	Post
% Unacceptable									
Soldier Method	16			16			.16		
Incident Method	22	21	23	17	21	15	.18	.19	.31
Task-HS Method	34	35	43	24	24	22	.15	.14	.22
Task-DPG Method	39	46	50	21	19	18	.29	.39	.47
Task-APG Method	41	38	48	22	18	20	.42	.32	.46
% Unacceptable or Marginal									
Soldier Method	36			22			.16		
Incident Method	29	28	28	20	19	17	.20	.19	.34
Task-HS Method	58	59	66	25	25	21	.12	.08	.09
Task-DPG Method	63	69	72	21	16	15	.28	.30	.40
Task-APG Method	63	61	69	22	20	17	.40	.38	.49
% Outstanding									
Soldier Method	15			17			.12		
Incident Method	18	19	17	18	18	17	.11	.11	.18
Task-HS Method	9	9	6	11	10	6	.13	.12	.28
Task-DPG Method	10	7	6	12	8	6	.18	.21	.26
Task-APG Method	11	12	7	13	14	7	.23	.22	.46

although also showing stricter standards for the Task-Based Methods, are not quite as striking. Between 7 and 12 percent are judged to be outstanding with the Task-Based Methods compared to 15 and 18 for the Soldier and Incident Methods.

A second conclusion that can be drawn from the means is that the Delphi sessions had very little effect on the means from the Incident Method (consistent with much prior research), but did have a significant effect on the means in from the Task-Based Methods. Unfortunately, the net effect of the Delphi sessions on the Task-Based results was to increase the strictness widening the gap between the Task-Based and other methods even further. This finding is somewhat at odds with other published research (e.g., Jaeger & Busch, 1984) which indicated that the means (standards) remain unaffected by the Delphi procedures which the variance decreases (i.e., agreement among judges increases).

There were differences between how Delphi sessions are typically conducted and how they were conducted in this research. In a typical delphi session, judgments are made independently and anonymously, pooled, summarized, and then fed back to the judges for another round of opinion. (Dalkey, 1969). This is somewhat different from the technique used in this research. As described in Chapter 2, initial judgments were made independently and anonymously, the judgments were pooled and the workshop leader choose for discussion the judgments in which there was the greatest disagreement. Participants were asked to explain their strategies aloud and the various strategies were discussed by the group.

One explanation for the shift in mean percent of soldiers performing in the post-Delphi task-based rating is that those who provided harsh ratings were more influential than those who provided lenient ratings.

It must be noted that the differences in the means and standard deviations are being attributed to the Delphi sessions when, in fact, since there was no control group, these results could reduce to other factors independent of Delphi, such as regression to the mean.

Differences in the standard deviations (across different judges) are a little difficult to interpret because there is some correlation between the means and standard deviations (as the mean percent moved away from 50, the standard deviations tended to decrease). In almost all cases, the standard deviations of the post-Delphi results were smaller than the pre-Delphi standard deviations (consistent with most prior research using Delphi).

There were also notable differences in the reliabilities associated with the different methods. The task-based methods, particularly those based on percent-Go score ratings, had significantly higher single-rater reliabilities than the other methods. This appears to be a result of stereotypical beliefs that 60 percent or 70 percent correct should be the minimum "passing" score.

Combining Multiple Standards

Much of the literature on standard setting concerns a single measure or a single dimension of performance. Project A and the Army Synthetic Validation Project, however, take a multidimensional perspective of job performance. A central issue to be considered when taking a multidimensional approach is the notion that an employee's job performance may be quite satisfactory in some areas but not satisfactory in others. Thus, decisions must be made regarding the extent to which more effective performance in some areas compensates for less effective performance in others. These decisions will dictate how standards for individual dimensions of performance should be combined into an overall performance standard.

The question of how to set an overall standard for job performance must necessarily be preceded by the development of a scale for assessing overall job performance. Several different approaches for developing such an overall performance scale, ranging from a simple linear composite to more complex conjoint measurements techniques, were examined as part of the Project A research (Sadacca, Park & White, 1986). A conjoint measurement approach (e.g., Luce & Tukey, 1964; Green & Srinivasan,

1978) asks judges to evaluate trade-offs among increments and decrements along different dimensions. For example, two soldiers, one having a slightly higher level of proficiency and a slightly lower level of motivation than the other, might be compared in terms of their overall contribution to the organization.

In its general form, the conjoint measurement model would not assume that the value of a performance increment is necessarily the same for different parts of different dimensions. It is possible, for example, that small decrements below minimum levels in some areas are balanced only by large increments above minimum levels in other areas. There are two special cases of interest in setting an overall performance standard. In the first case, no amount of increment in other areas can compensate for below standard performance on any other dimension. Using this model, known as the Multiple Hurdles Model, an examinee fails the overall standard if he or she fails any of the individual standards. The other special case of interest is a strictly linear model, when overall performance is measured by a weighted sum of the individual performance measures. Using this model, known as the Compensatory Model, a decrement in one performance area could be compensated for by an equal increment in another area.

The conjoint measurement approach attempts to mathematically model the *qualitative laws* that judges use to combine information and make judgments. The advantage of this approach is that it allows for nonlinear variations of the multiple-hurdles and compensatory models to be discovered. The disadvantage of the approach is that it relies on the ability of the judges to combine multi-source information in order to make judgments. The general procedure is as follows: Judges are provided with information about performance standards on several job dimensions and are asked to combine this information and provide an overall job performance standard. A mathematical model is then constructed to capture the judges' policy. This model is then used to transform information on individual dimensions into an overall score that can be compared with the overall standard.

This section describes results of analyses in which the conjoint measurement approach to setting overall job performance standards was applied to data collected from the seven Phase II MOS. The purpose of this research was to develop mathematical models of the strategies that judges used when combining standards on individual job dimensions into an overall standard. The number of individual dimensions which were combined in each MOS ranged from three to five. Note that these dimensions were the same as those presented in the other standard setting exercises. The number of judges correspond to the sample sizes reported in Table 2-2 on page 2-4.

Procedure

A conjoint measurement approach was used to determine how the judges evaluated the trade-offs between different increments and decrements of performance on different dimensions when setting overall performance standards. Within each MOS, subjects were provided information on the same 64 hypothetical soldiers that varied in their performance on standard setting dimensions. Depending on the MOS, three to five dimensions were used. The soldiers' performance on each dimension was described as Unacceptable (U), Marginal (M), Acceptable (A), or Outstanding (O). For example, the performance of a given hypothetical soldier may have been described as "Unacceptable" on two particular dimensions and "Acceptable" on a third. The judges were asked to provide an overall performance rating (Rating Scale: U - Unacceptable, M - Marginal, A - Acceptable, O - Outstanding) for each of the 64 hypothetical soldiers.

Results

All ratings of overall performance and descriptions of performance on individual performance dimensions were converted according to a four-point integer scale (such that U=0, M=1, A=2, and O=3). A regression equation was then computed for each MOS using the mean rating of overall performance for each of the 64 hypothetical soldiers across all judges and the integer scaling of each individual dimension. Tables 8.18-8.20 contain results of these regressions.

The intercept, raw regression coefficients, and percentage of variance in overall ratings accounted for (R^2) associated with the regression equation computed for each MOS are reported in Table 8.18. Examination of these results leads to several observations. First, the linear model appears to do a good job of accounting for variance in the overall ratings of performance. This is indicated by the relatively high R^2's reported in the column furthest to the right. These values range from .43 to .63, with an average percentage of variance accounted for equal to .51.

A second observation concerns the variability in the size of the regression coefficients within the equations for several of the MOS. For example, in the regression equation computed for 94B, the regression weight for dimension 23 (Food Preparation) was approximately twice as large as the regression weights for dimension 11 (Pack and Load) and dimension 13 (Operate/Assemble/Install). Findings such as this suggest that performance on each of the individual dimensions did not contribute equally to the ratings of overall performance. Instead, performance on some dimensions was more influential than performance on other dimensions.

Note, however, that caution must be observed in the interpretation of raw regression coefficients with respect to the relative influence of independent variables on a given criterion. For one reason, if two variables have the exact same relationship with a particular criterion, but do not have the exact same variance, then the variable with the greater variance will receive the smaller raw regression weight. Table 8.19 reports the standardized regression coefficients which correspond to the raw coefficients in Table 8.18. The relative differences among coefficients within these standardized equations are approximately the same as those reported above. This indicates that the variability among the raw coefficients was not due to differences in the variances across the individual performance dimensions.

TABLE 8.18

OVERALL STANDARD SETTING
REGRESSION RESULTS BY MOS AND DIMENSION

Military Occupational Specialty	Intercept	Dim 2	Dim 4	Dim 5	Dim 7	Dim 8	Dim 9	Dim 11	Dim 13	Dim 15	Dim 16	Dim 17	Dim 18	Dim 19	Dim 22	Dim 23	R2
16S:	-.50	.24			.27			.20		.22			.18				.50
19K:	-.51	.26			.22			.18		.26			.18				.50
67N:	.55					.42			.01*			.25					.43
76Y:	-.09							.16			.22	.36		.25			.54
88M:	-.23		.17				.18	.17		.41		.15					.49
91A:	-.47			.29								.13	.16	.15	.36		.51
94B:	-.01*							.23	.25							.48	.63

*Coefficient is not significant, p > .05.

Dimension 01: Operate/Maintain Individual Weapons
Dimension 02: Operate Crew-served Weapons
Dimension 03: Tactical Movements/Reconnaissance
Dimension 04: Navigate
Dimension 05: Administer First-Aid/NBC
Dimension 06: Perform Under Adverse Conditions
Dimension 07: Detect/Identify Targets
Dimension 08: Inspect/Repair/Maintain Mech Sys
Dimension 09: Inspect/Repair/Electrical Systems
Dimension 10: Use Technical References
Dimension 11: Pack and Load
Dimension 12: Construct/Assemble/Install

Dimension 13: Operate/Assemble/Install
Dimension 14: Operate Electronic Equipment
Dimension 15: Operate Vehicles/Heavy Equipment
Dimension 16: Operate Keyboard/Type
Dimension 17: Administration/Records Keeping
Dimension 18: Oral Communication
Dimension 19: Written Communication
Dimension 20: Analyze/Interpret Information
Dimension 21: Draw/Sketch
Dimension 22: Provide Medical Treatment
Dimension 23: Food Preparation
Dimension 24: Demonstrate Leadership

TABLE 8.19

OVERALL STANDARD SETTING REGRESSION RESULTS BY MOS AND DIMENSION (STANDARDIZED REGRESSION COEFFICIENTS)

Military Occupational Specialty	Intercept	Dim 2	Dim 4	Dim 5	Dim 7	Dim 8	Dim 9	Dim 11	Dim 13	Dim 15	Dim 16	Dim 17	Dim 18	Dim 19	Dim 22	Dim 23	R2
16S:	.0	.34			.38			.28		.31			.26				.50
19K:	.0	.37			.32			.25		.37			.26				.50
67N:	.0					.57			.01			.47					.43
76Y:	.0							.23			.31	.51		.36			.54
88M:	.0		.23				.24	.23		.55		.20					.49
91A:	.0			.39								.18	.22	.21	.49		.51
94B:	.0							.31	.34							.65	.63

Dimension 01:	Operate/Maintain Individual Weapons
Dimension 02:	Operate Crew-served Weapons
Dimension 03:	Tactical Movements/Reconnaissance
Dimension 04:	Navigate
Dimension 05:	Administer First-Aid/NBC
Dimension 06:	Perform Under Adverse Conditions
Dimension 07:	Detect/Identify Targets
Dimension 08:	Inspect/Repair/Maintain Mech Sys
Dimension 09:	Inspect/Repair/Electrical Systems
Dimension 10:	Use Technical References
Dimension 11:	Pack and Load
Dimension 12:	Construct/Assemble/Install
Dimension 13:	Operate/Assemble/Install
Dimension 14:	Operate Electronic Equipment
Dimension 15:	Operate Vehicles/Heavy Equipment
Dimension 16:	Operate Keyboard/Type
Dimension 17:	Administration/Records Keeping
Dimension 18:	Oral Communication
Dimension 19:	Written Communication
Dimension 20:	Analyze/Interpret Information
Dimension 21:	Draw/Sketch
Dimension 22:	Provide Medical Treatment
Dimension 23:	Food Preparation
Dimension 24:	Demonstrate Leadership

TABLE 8.20

OVERALL STANDARD SETTING
REGRESSION RESULTS BY MOS AND DIMENSION
(VARIANCE ATTRIBUTED TO EACH DIMENSION

Military Occupational Specialty	Dim 2	Dim 4	Dim 5	Dim 7	Dim 8	Dim 9	Dim 11	Dim 13	Dim 15	Dim 16	Dim 17	Dim 18	Dim 19	Dim 22	Dim 23	R2
16S:	.11			.14			.08		.10			.07				.50
19K:	.14			.10			.08		.13			.05				.50
67N:					.32			.00			.11					.43
76Y:							.05			.10	.26		.13			.53
88M:		.05				.08	.05		.30		.04					.49
91A:			.15								.03	.05	.04	.24		.51
94B:							.10	.11							.42	.63

Dimension 01: Operate/Maintain Individual Weapons
Dimension 02: Operate Crew-served Weapons
Dimension 03: Tactical Movements/Reconnaissance
Dimension 04: Navigate
Dimension 05: Administer First-Aid/NBC
Dimension 06: Perform Under Adverse Conditions
Dimension 07: Detect/Identify Targets
Dimension 08: Inspect/Repair/Maintain Mech Sys
Dimension 09: Inspect/Repair/Electrical Systems
Dimension 10: Use Technical References
Dimension 11: Pack and Load
Dimension 12: Construct/Assemble/Install

Dimension 13: Operate/Assemble/Install
Dimension 14: Operate Electronic Equipment
Dimension 15: Operate Vehicles/Heavy Equipment
Dimension 16: Operate Keyboard/Type
Dimension 17: Administration/Records Keeping
Dimension 18: Oral Communication
Dimension 19: Written Communication
Dimension 20: Analyze/Interpret Information
Dimension 21: Draw/Sketch
Dimension 22: Provide Medical Treatment
Dimension 23: Food Preparation
Dimension 24: Demonstrate Leadership

A second concern regarding the interpretation of regression coefficients with respect to the relative influence of independent variables on a given criterion is the effects of multicollinearity (i.e., covariation among the independent variables). When multicollinearity exists, two independent variables may both be more highly related to a given criterion variable than is a third independent variable, yet the size of their respective regression coefficients may both be smaller than that of the third variable. This could occur, for instance, if the first two variables were highly correlated with each other but uncorrelated with the third. The results in Table 8.20, however, indicate that, as designed, the individual dimensions associated with each MOS were essentially orthogonal to one another for the 64 hypothetical soldiers being rated. This table reports the percent of variance in the ratings of overall performance within each MOS accounted for uniquely by each individual performance dimension. The fact that the sum of these percentages within each MOS is approximately equal to the corresponding R^2 shows that practically none of the variance in overall ratings is accounted for jointly by two or more of the individual dimensions.

One final observation regarding the results reported in Table 8.18 concerns the intercepts of the seven regression equations. With the exceptions of those associated with the regression equations computed for 67N (intercept = .55) and 94B (intercept = -.01), all of the intercepts are significantly negative. These negative intercept values indicate that the overall ratings associated with the corresponding MOS were lower than the weighted average of performance on the individual dimensions. This suggests that the judges in those MOS did not use a fully compensatory rating policy (whereby low performance on one or more dimensions is offset by high performance on one or more others). Specifically, the negative intercept indicates that, on average, poor performance is not counterbalanced by good performance. Instead, ratings of overall performance appear to have been disproportionately affected by low performance on individual dimensions, suggesting that judges in these MOS used a rating policy representing a compromise between those indicated by the compensatory and multiple hurdle models.

CHAPTER 9

SUMMARY AND CONCLUSIONS FOR THE STANDARD SETTING ANALYSES

Lauress L. Wise and Deborah L. Whetzel (AIR)

Research Questions Revisited

We return to the standard setting research questions posed in Chapter 1. For each question we summarize the conclusions that we draw from the analyses described in Chapter 8. First, the following questions were raised concerning the individual standard setting methods:

(1) For each instrument, to what extent did different types of judges (NCO vs Officer, FORSCOM vs TRADOC) differ in terms of the mean levels of the standards that they set or the level of agreement (as measured by the standard deviation of the judgments across judges or by reliability estimates)?

Most of the evidence indicated a high level of similarity across judge types in both mean levels and the degree of agreement. There were some instances where greater agreement or slightly different standards were produced by Officers and by FORSCOM judges. Such differences were, however, small in comparison to the very great differences among the different methods.

(2) For the Critical Incident and the Task instruments, were the post-Delphi judgments significantly different from the initial judgments in terms of means and agreement levels?

The Delphi sessions did have a very significant impact on the degree of agreement among judges (with significantly higher consistency in the post-Delphi sessions) and, in some cases, in the overall levels of the standards that were set. For the task-based methods, the Delphi sessions led to even stricter standards; for the Incident

Method, the Delphi sessions produced greater agreement but no significant shift in mean level.

(3) For the Task instrument, were there differences (in mean levels, agreement levels, and Delphi changes) among standards based on the hypothetical soldier ratings, the detailed information percent-go ratings, and the abbreviated information percent-go ratings?

The hypothetical soldier ratings did lead to somewhat less severe standards, but also produced less agreement among judges. There were only small overall differences between the Detailed Percent-Go and the Abbreviated Percent-Go results. Providing specific score sheets (the Detailed Method) actually led to slightly lower reliabilities and to differences in severity that were small and not consistent across performance levels and across Delphi conditions.

(4) Were there differences among the different instruments (and the three different approaches within the task-based instrument) in terms of means, agreement levels, Delphi effects, and discrepancies across judge types?

The Incident Method produced standards that matched the judges direct estimate of the percent of soldiers performing at each level (22% unacceptable compared to the direct estimate of 16%; 29 percent less than acceptable compared to the direct estimate of 36%; and 18% outstanding compared to the direct estimate of 15%). By comparison, the Task Methods led to standards such that the percent unacceptable was 35% or more, the percent less than acceptable was about 60%, and the percent outstanding was 10% or less. The Delphi sessions improved the reliabilities of the Incident Method judgments (from .19 to .31) without changing the mean levels of the standards. The Delphi sessions led to similar reliability increases for the Task Method, but at the expense of significant changes in the mean standards in the wrong direction (increase severity with larger discrepancies from the direct judgments).

Second, for the exercise on combining multiple standards, the basic questions for analysis were:

(5) To what extent did a compensatory model explain the judges ratings better than a multiple hurdles model?

The results were highly consistent with a general compensatory model.

(6) Did the judges give equal weight to each performance dimension?

No. The findings suggest that performance on each of the individual dimensions did not contribute equally to the ratings of overall performance. For instance, judges in 94B appear to have placed twice as much emphasis on dimension 11 (Food Preparation) as they did on dimension 13 (Operate/Assemble/Install).

(7) Were the overall ratings significantly higher or lower than the simple average?

For most of the MOS, the overall ratings were lower than the simple average of performance across the individual dimensions. This finding suggests that ratings of overall performance were disproportionately affected by low performance on individual dimensions.

Recommendations

The Incident Method led to good agreement with the Soldier Method results. While the reliabilities were only modest, they were significantly improved by the Delphi process. For Phase III, we will revise the dimension descriptions working toward greater agreement with the performance dimensions used for job description. We will examine the statistics for each incident, and incidents about which there was significant disagreement (high standard deviations) will be replaced insofar as possible. We also will check the mean effectiveness levels (from the retranslation workshops) against the mean ratings (from the Phase II workshops) and adjust or eliminate items where there is significant disparity. Particular attention also will be paid to modifying or replacing items that describe behaviors specific to a single MOS.

The Task methods led to very good agreement, but unrealistically stringent standards when compared to actual performance distributions from Project A. We propose one more attempt to introduce normative information into the rating and Delphi sessions as a means of encouraging more moderate standards. The hypothetical soldier ratings were slightly more realistic in comparison to the percent-go rating methods, but at the price of significantly less reliability. Otherwise, the detailed methods did not have much to commend them. For Phase III, we also will need more explicit procedures for having each group of judges substitute MOS-specific tasks for sample tasks that are not appropriate to their MOS.

REFERENCES

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1987). Analysis of criterion - measures: The modeling of performance. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta.

Chia, W. J., Hoffman, R. G., Campbell, J. P., Szenas, P. L., & Crafts, J. L. (1989). Analysis of job components: The development and evaluation of alternative methods. In L. L. Wise, J. M. Arabian, W. J. Chia, & P. L. Szenas (Eds.), Army Synthetic Validity Project: Report of Phase I results. (ARI Technical Report 845) Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A 219-926)

Claudy, J. (1978). Multiple regression and validity estimation in one sample. Applied Psychological Measurement, 2, 595-607.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York: Wiley.

Dalkey, N. (1969). The Delphi method: An experimental study of group opinions. CA: Rand corp.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. Journal of Consumer Research, 5, 103-123.

Jaeger, R. M., & Busch, J. C. (1984). The effects of a delphi modification of the Angoff-Jaeger standard-setting procedure on standards recommended for the National Teacher Examinations. Paper presented at the joint annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA: (ERIC Document 246 091).

Kane, M. T., Kingsbury, C., Colton, D., & Estes, C. (1989). Combining data on criticality and frequency in developing test plans for licensure and certification examinations. Journal of Educational Measurement, 26, 17-27.

Lord, F. M. & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. New York: Addison-Wesley.

Luce, R., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. Journal of Mathematical Psychology, 1, 1-27.

Owens-Kurtz, C. K., & Peterson, N. G. (1989). Development of an attribute taxonomy and its application in the formation of synthetic validity composites. In L. L. Wise, J. M. Arabian, W. J. Chia, & P. L. Szenas, P.L. (eds.), Army Synthetic Validity Project: Report of Phase I results. (ARI Technical Report 845) Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A 219-926)

Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. L., Houston, J. S., & Toquam, J. L. (1987, April). Identification of predictor constructs and development of new selection/classification tests. Paper presented to the Second Annual Conference of the Society of Industrial and Organizational Psychology, Atlanta.

Peterson, N. G., Owens-Kurtz, C. K., Hoffman, R. G., Arabian, J. M. & Whetzel, D. L. (In preparation). Army Synthetic Validation Project: Appendices to Report of Phase II results. (ARI Technical Report 845) Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A 219-926)

Peterson, N. G., Rosse, R. R., & Owens-Kurtz, C. K. (1989). Applications and evaluation of synthetic methods of forming predictor composites for Army jobs. In L. L. Wise, J. M. Arabian, W. J. Chia, & P. L. Szenas (Eds.) Army Synthetic Validity Project: Report of Phase I results. (ARI Technical Report 845) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A 219-926)

Sadacca, R. A., Park, M. V., & White, L. (1986). Weighting performance constructs in composite measures of job performance. Paper presented at the Annual Meeting of the American Psychological Association, Washington, D.C.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Szenas, P. L., & McHenry, J. J. (1989). Exploring the relationships among rater characteristics and fidelity of job description judgments. In L. L. Wise, J. M. Arabian, W. J. Chia, & P. L. Szenas (Eds.), Army Synthetic Validity Project: Report of Phase I results. (ARI Technical Report 845) Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A 219-926)

Wilkinson, L. (1988). SYSTAT: The system for statistics. Evanston, IL: SYSTAT, Inc.

Wing, H., Peterson, N. G., & Hoffman, R. G. (1984). Expert judgments of predictor-criterion validity relationships. Paper presented at the annual convention of the American Psychological Association, Toronto.

se, L. L., Campbell, J. P., McHenry, J. J., & Hanser, L. R. (1986, August). A latent structure model of job performance factors. Paper presented at the 85th American Psychological Association Convention, Washington, DC.

se, L. L., Campbell, J. P., & Peterson, N. P. (1987, April). Identifying optimal predictor composites and testing for generalizability across jobs and performance constructs. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta.

se, L. L., Peterson, N. G., Rosse, R. R., & Campbell, J. P. (1989). Comparative analyses of empirical and synthetic job performance prediction equations. In L. L. Wise, J. M. Arabian, W. J. Chia, & P. L. Szenas (Eds.), Army Synthetic Validity Project: Report of Phase I results. (ARI Technical Report 845) Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A 219-926)

se, L. L., Peterson, N. G., Hoffman, R. G., & Arabian, J. M. (In preparation). Army Synthetic Validation Project: Phase II instruments. (ARI Research Note). Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences.